A general proposal for model-free difference-in-differences

Daniel J. Henderson*[1,2,3] and Stefan Sperlich†[4]

[1]Department of Economics, Finance and Legal Studies, University of Alabama
[2]School of Mathematical Sciences, Nankai University
[3]Institute for the Study of Labor (IZA)
[4]Geneva School of Economics and Management, University of Geneva

November 17, 2021

Abstract

We propose a general framework for model-free difference-in-differences analysis with confounders. Following the natural steps in practice, we start by searching for the preferred data setup, namely the simultaneous selection of confounders and potential data (outcome) transformations. We then offer a test for the credibility of identification assumptions. The treatment effects themselves are estimated in two steps: first, the heterogeneous effects stratified along the confounders, then second, the average treatment effect(s) for the population(s) of interest. We suggest bootstrap procedures to calculate the standard errors of these estimates, as well as for significance tests. We study the asymptotic statistics as well as the finite sample behavior (via simulations) of our tests and estimators. We address practical issues that arise such as bandwidth selection, incorporating sample weights and dealing with discrete data in both the outcome variable and set of confounders. These are addressed in a setting whereby we look at the impact of the Deferred Action for Childhood Arrivals (DACA) program on human capital responses of non-citizen immigrants. We find that past (linear parametric) estimates are similar regarding (albeit underestimating) the positive impact of DACA on school attendance (for individuals aged 14-18) and the positive impact on high school completion, but at the same time, fail to identify the negative impact of DACA on school attendance for college aged individuals.

Keywords: causal estimation, difference-in-differences, kernel, nonparametric, treatment effects
JEL classification codes: C14, C21, I21, J15

*Department of Economics, Finance and Legal Studies, University of Alabama, Tuscaloosa, 35487-0224, USA; djhender@cba.ua.edu

†Geneva School of Economics and Management, University of Geneva, 1211 Geneva 4, Switzerland; stefan.sperlich@unige.ch

1 Introduction

Arguably the most popular estimation technique for causal analysis is the difference-in-differences (DiD) approach. This is feasible when a short panel or cohorts (repeated cross-sections) of individuals are observed both before and after an intervention has taken place.[1] We call such political intervention or similar event a 'treatment'. Although the basic concept for identifying the causal effect applies to more complex situations [Lechner, 2011], we limit our considerations to the case of a single treatment and two groups (i.e., the treatment group, $D = 1$ and the control group, $D = 0$). The primary assumption behind this method for identifying the treatment effect on the treated is that without such intervention, the outcome variable of interest Y experienced in both groups would have developed similarly over time. This is also known as the 'common trend' or 'parallel path' condition. Intuitively, this insinuates that there is a constant difference between the two groups, disturbed only by this particular treatment.

In many cases, it is unlikely that this difference is independent of other factors (e.g., 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 Y, while neither originate from the treatment itself. In the former example, we think of a possible interaction between a (pre-)condition and time, and in the latter, an exogenous change of conditions over time. Both of these fears can be mitigated by proper conditioning, say by considering DiD with confounders X. While (for identification) a common trend (conditional or unconditional) is only required for a given period before and after the treatment, it seems reasonable to assume that this should also hold for a short/long period before the intervention.[2]

For all of the considerations above, we focus our attention on the DiD statistic:

$$\{E[Y_t|X_t, D = 1] - E[Y_t|X_t, D = 0]\} - \{E[Y_{t-1}|X_{t-1}, D = 1] - E[Y_{t-1}|X_{t-1}, D = 0]\}, \tag{1}$$

where $E[Y|X, D]$ is the expectation of the outcome Y conditional on the set of confounders X and treatment status D in time period t. When the treatment takes place between periods $t - 1$ and t, this statistic gives the conditional treatment effect on the treated (from which you can obtain average effects).[3] This is based on the assumption that (1) had been zero for all $X = x$ without treatment. To identify a causal effect, work with a scale for Y and a set X such that (1) is zero for periods without treatment. This statistic turns a bane into a boon: while it may be difficult to convince others that essential identifying assumptions are fulfilled, an appropriate statistic can guide you. Assuming data is available in at least one additional period prior to the treatment (i.e., period $t = -1$), we can check if (1) is zero for all X in periods prior to the treatment (i.e., the development between periods $t = -1$ and $t = 0$). While it is true that this is not the identification condition needed, it would at least empirically support its credibility.

Equation (1) is far more useful than simply being used to estimate an average treatment effect on the treated. We therefore study its estimation (i.e., heterogenous treatment effect on the treated), sample average (i.e., average treatment effect on the treated), and the analogue of its squares (i.e., test statistics). In each case, we will study the asymptotic and finite sample properties. In practice,

[1] Our methods are designed for repeated cross-sections. These can be simplified to the case of balanced panels.

[2] The same can be said about periods after the treatment. However, this only holds for treatments that simply shift the development of Y by a constant, a somewhat strong hypothesis.

[3] For simplicity, we will consider three time periods $t = -1, 0$ and 1. The treatment will occur between periods 0 and 1, while period $t = -1$ will be used for pre-treatment analysis.

it is likely preferable to rely on resampling methods than estimates of these complex asymptotics. We therefore introduce, in parallel, bootstrap methods to approximate standard errors of estimates and p-values for our test statistics. This is a challenge as it requires that our bootstrap procedures generate data under complex null hypotheses.

Without confounders, the linear DiD estimator is identical to the nonparametric average treatment effect on the treated estimator. However, in the presence of confounders, this need not be the case [Meyer, 1995]. Nonparametric estimation is often avoided for fear of the curse of dimensionality. While this curse 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 can be estimated at the parametric (i.e., root-n) rate without asymptotic bias. Unless the number of variables increases with the sample size, only continuous confounders count for the curse. Moreover, if the unconditional treatment effect on the treated is of interest, you typically need to have more than three continuous variables to be affected asymptotically. Many economic variables are discrete (e.g., union status) and many continuous variables are measured discretely (e.g., years of education). A nonparametric approach is reasonable, even computationally, when all confounders are discretely measured.[4]

This is not an uncommon phenomenon. For example, solely looking at the *American Economic Journal: Applied Economics*, examples include Ang [2019], who looked at the impact of the Supreme Court (in 2013) striking down parts of the Voting Rights Act on long-run voter turnout. His model regressed 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. Additional regressions were separately estimated for whites and non-whites to allow for coefficients to vary across groups. Panhans [2019, pp. 24-25] looks for adverse selection in the Affordable Care Act health insurance exchanges. A supplemental section of his paper uses difference-in-differences with a set of fixed effects which are not exhaustive and hence are not identical to nonparametric estimates. McKenzie et al. [2014, pp. 68] 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. Kuka et al. [2020] examine human capital responses to the availablility of the Deferred Action for Childhood Arrivals (DACA) program. In this paper, in addition to having all binary right-hand-side variables, their outcome variables are binary. Authors usually have a mix of discrete and continuous variables. We consider this general setting and argue that empirical researchers should be more concerned about systematic biases and inconsistency due to model specification than the curse of dimensionality in model-free estimation.

The contribution of this article is the introduction of a general, model-free DiD based causal analysis under the potential presence of confounders. We start by presenting a procedure to find the scale of Y and the set of confounders X that allow for identification of treatment effects, exhibiting the wanted 'parallel trend'. We then estimate the identified conditional and unconditional treatment effects on the

[4]It is relatively straightforward to employ parametric or semiparametric versions of our estimators/tests. We simply replace the conditional expectations with their parametric or semiparametric counterparts (c.f., Abadie [2005]). However, these strict parametric assumptions may or may not be validated by economic theory and misspecification of functional form typically leads to biased and inconsistent estimates.

treated. The procedure is concluded by the introduction of nonparametric tests for significant treatment effects for which the multivariate test for significance of all conditional treatment effects equals the one for testing credibility of the 'parallel trend' assumption which was applied in a previous time period (i.e., $t = -1$ and 0).

We present a set of simulations to confirm our asymptotic developments. As it is uncommon for nonparametric estimators to be estimated at parametric rates,[5] it is particularly interesting to see their performance with very small samples. The performance of our covariate selector, estimators and tests perform admirably, even in these small sample settings.

To highlight the usefulness of our approach in an empirical setting, we re-examine the results of Kuka et al. [2020]. We find mixed evidence that their set of confounders satisfy the 'parallel trend' assumption. Regarding their treatment effects estimates, their models underestimate (albeit not statistically) 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). Further, they fail to identify the negative impact of DACA on school attendance of college aged individuals (19-22). With respect to enrolling in college, as is in their paper, we find that these effects are insignificant.

The remainder of our paper proceeds as follows: Section 2 presents the basics of model-free conditional DiD analysis. Section 3 develops a scale and variable selection tool, including a formal test for checking the validity of the 'parallel path' assumption. Section 4 presents the estimator and its asymptotic properties together with their bootstrap approximations. This is followed by a general format for nonparametric significance tests with bootstrap procedures in Section 5. Section 6 provides our simulations which confirm our theory, while Section 7 contains our application. Section 8 discusses various practical issues needed for implementation of our approach such as data driven bandwidth selection, incorporating sample weights, and an algorithm which includes `R` functions to apply said procedures. Section 9 concludes and provides directions for future research. All theoretical proofs and details on the `R` procedure code can be found in the Supplemental Appendix.

2 Nonparametric difference-in-differences

We begin this section by first recalling ideas, definitions and assumptions of DiD with confounders, suppressing the DiD approach without them for sake of brevity. For more details on both you may consult the compendium of Frölich and Sperlich [2019]. For a linear parametric DiD with confounders see Słoczyński [2018] or Sant'Anna and Zhao [2020] for its so-called double robust version (i.e., propensity score weighting and regression). In a fully parametric context, you get a consistent estimator of the treatment effect, if either the propensity score or the regression function is correctly specified, and you don't if both are not. In nonparametric estimation, there is no role for it as both functions are 'correctly specified', and doing both will not improve efficiency.[6] Abadie [2005] proposed a DiD with nonparametric propensity score weighting. As there is no general superiority of one approach over the other, we stick with our equation (1) and thereby to nonparametric regression. The practical advantage

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

[6] Using series estimators, the approximation bias can (easily) be substantial. Again, this does not apply to our setting. We do not advise using them. We are unaware of any paper supporting the guess that double robust estimators are more efficient when both (propensity score and regression function) are poorly approximated.

is that we do not need to jump between our steps of analysis from nonparametric propensity estimation to nonparametric regression and back. We also avoid numerical problems that occur when dividing by nonparametric estimates of (potentially) small or tiny propensities.

Before we proceed, we should add/expand our thoughts on conditioning variables. Thus far, we have only looked at them as classical confounders (i.e., variables that (partly) predict D and Y). As Frölich and Sperlich [2019] discuss, there are at least two additional reasons someone might condition on certain covariates. One is to measure a direct or 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, and whose inclusion can improve the statistical analysis by resting noise. A further reason can be to study certain heterogeneities of treatment effects. Which covariates to choose is seemingly the researcher's choice, but this has implications for the assumptions below. As we condition on both, confounders and those additional covariates, we will henceforth speak of 'covariates' in general, not solely confounders. Further, we use the notation thinking of cohorts where T stands for the random variable indicating the time period. Where appropriate we will discuss the case of panel data explicitly.

2.1 Difference-in-differences with covariates

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. For estimating the treatment effect, we do not have to *a priori* know and observe the exact conditions, say $\tilde{X}$ under which this holds; it is sufficient to observe a set of indicators X, and know the scale of Y, such that stochastically speaking, this holds in the mean. We call X the set of covariates, and will extend it where appropriate further below. 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. As usual, observations at $t = 0$ are supposed to be free of anticipation effects; else you only measure the treatment minus anticipation effect. Considering what we have said above, we first assume

Assumption 1 For a set of observed covariates (X) that are not affected by the treatment (D), the difference in potential outcomes under no treatment (Y^0) between the treatment and control group is the same before ($t = 0$) and after ($t = 1$) treatment:

$$\{E\left[Y^0_{t=1}|x, D = 1\right] - E\left[Y^0_{t=1}|x, D = 0\right]\} \quad = \quad \{E\left[Y^0_{t=0}|x, D = 1\right] - E\left[Y^0_{t=0}|x, D = 0\right]\}. \tag{2}$$

It is important to note here that you 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) highlights the link to matching estimators based on a conditional comparison of treatment versus control groups after treatment ($t = 1$). In that setting, we assume that the vector X accounts for all differences in Y^0 such that the left-hand-side of (2) is zero, and if not, its average over all x is the bias of the average treatment for the treated matching estimator.

In the DiD approach, we only assume that this difference is the same before the treatment, such that we can use pre-treatment data for bias correction. Therefore, calling Assumption 1 'bias stability' is perhaps more appropriate as it does not deceptively insinuate a parallel path of Y^0. In addition to Assumption 1, we need

Assumption 2 A common support condition (CSC), whereby

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

$\forall x \in \mathcal{X}, \forall (t, d) \in \{(0, 0), (1, 0), (0, 1)\}$, with $\mathcal{X}$ being the domain of x.

Loosely speaking, we require that X takes similar values for each group in each time period. There should be no value of X whereby we cannot find a counterfactual match. This says little about the underlying distribution of X within each group in each time period. These assumptions allow for the inclusion of covariates changing over time or even having a time trend.

While Assumption 2 can be checked, Assumption 1 is the usual 'non-testable identification condition'. However, as said in the introduction, Assumption 1 is not very credible if it does not hold at least one period before as well. Consequently, we can apply both assumptions to periods prior to treatment ($t = -1$ and 0) and use data from those periods to verify Assumption 1, which is feasible because for $t < 1$, $Y_t^0 = Y_t$.

Let us denote the conditional expectations for each year and treatment group by

$$m_{dt}(x) = E[Y|X = x, D = d] \ , \quad d = 0, 1, \ t = -1, 0, 1 \ . \tag{3}$$

Under Assumptions 1 and 2, the conditional average treatment effect on the treated (TT) for a given x is identified by

$$TT_x = \{m_{11}(x) - m_{01}(x)\} - \{m_{10}(x) - m_{00}(x)\}, \tag{4}$$

and consequently, the unconditional treatment effect on the treated, 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 group $D = T = 1$. We will also comment on the TT that results from integrating over all individuals with $D = 1$ (TT_b).

For balanced panels, TT_a and TT_b are the same. We speak of TT when we refer to both. Again, we do not require a balanced panel. We therefore may not be able to observe X for all people at $t = 0$ or $t = -1$; the X for individual i may only be observed for the time point t when the individual's outcome is observed (Y_i). All our methods and results are applicable to the simpler case of balanced panels. Unfortunately, assuming a balanced panel from the onset does not lead to equivalent results for repeated cross-sections, but it does typically simplify the asymptotics.

2.2 Nonparametric conditional expectations

In practice, most academic 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 one had only discrete X which we would decompose into dummies; a saturated linear model would require to include all these dummies together with all their interactions of any order.[7] We are not aware of any practical work having done this; such inclusion

[7] The common practice of splitting the sample to obtain heterogenous estimates in the parametric world is valid assuming the functional form assumed is correctly specified and that there is a sufficient number of observations in each estimation. This practice addresses parameter heterogeneity, it does not cure functional form misspecification.

is usually arbitrary, guided by numerical convenience. Clearly, if just one covariate is continuous or discrete with many values, this problem is heavily aggravated. Nonparametric methods remove these concerns and let the data pick the appropriate variables and interactions. Researchers often ignore the use of these methods and use the "curse of dimensionality" as their argument against them. However, 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 the DiD compared to all treatment competitors when having nonexperimental data, as the differencing already accounts for many counfounders.

Suppose the scale of Y and the set of covariates are given. Then, in a first step, for each group d and each time period t (with notation T for treating time as a random variable), we can estimate their mean functions $m_{dt}(x)$ from the data set $\{Y_{it}, X_{it}\}_{i=1}^{n_{dt}} | D_{it} = d$. For sake of notation, the vector of covariates X_{it} is split 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 (discrete) variables $X_{it}^c = (X_{it,1}^c, ..., X_{it,k}^c)$, respectively. 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 $d_{X_i,x} := \sum_{l=1}^{k} 1\{X_{it,l}^c \neq x_l^c\}$ 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 and λ are our smoothing (bandwidth) parameters.[8] Under some standard regularization conditions outlined in Racine and Li [2004], namely on the smoothness of $m_{dt}(\cdot)$ and density $f(\cdot)$ of X^s, for $\lambda, h \to 0$ when $n_{dt} \to \infty$, we have

$$\sqrt{n_{dt}h^p}\left\{\widehat{m}_{dt}(x) - m_{dt}(x) - B_{dt}(x, h, \lambda)\right\} \to N(0, \Omega_{dt}(x)) \tag{5}$$

where the conditional mean estimator, given by

$$\widehat{m}_{dt}(x) = \frac{\sum_{i=1}^{n_{dt}} K(X_i, x, h, \lambda) Y_i}{\sum_{i=1}^{n_{dt}}, K(X_i, x, h, \lambda)} \tag{6}$$

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 it is more efficient in practice. The well established bias equals

$$\begin{aligned} B_{dt}(x, h, \lambda) &= h^2 \left[\nabla^t m_{dt}(x) \nabla f(x)/f(x) + tr\{\nabla^2 m_{dt}(x)\}\right] \int w(u) u^2 du \\ &\quad + \lambda \sum_{\tilde{x}, d_{\tilde{x},x}=1} \left\{m_{dt}(x^s, \tilde{x}^c) - m_{dt}(x)\right\} f(x^s, \tilde{x}^c) f^{-1}(x) \end{aligned} \tag{7}$$

$$\text{and} \quad \Omega_{dt}(x) = Var(Y|x, D = d, T = t) \int w^2(v) dv \; f^{-1}(x), \tag{8}$$

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. Equation (5) shows only the number of continuous covariates (p) impedes the parametric rate (root-n) of convergence. By using local-polynomials, we could achieve a faster rate for the bias (h^2) as long as we are willing to accept

[8] In practice, we use a separate bandwidth (h_l, λ_k) for each covariate. As is common in the theoretical literature, to avoid the notational burden, we treat them as equal in our theory ($h_l \; \forall l$, and $\lambda_k = \lambda \; \forall k$). The theoretical extension is straightforward.

higher smoothness conditions on $m_{dt}(\cdot)$ and the densities of the continuous covariates.[9] Although this is standard in the semiparametric econometrics literature, especially in the context of sieve estimators, we abstain from theoretical tricks and concentrate on practical procedures.

3 Selection of covariates and scale

Before estimating the TT, we first need to decide on the set of covariates, and the scale of the response Y. While economic theory or intuition may tell you assuming a common trend is sensible, it does not necessarily tell you the right scale nor the right set of conditioning variables $X^S \subseteq X$ for which it holds numerically. What is often criticized as the bane of DiD estimation, we suggest here to turn into a boon. We can use economic logic to help specify the causality model, but allow the procedure to determine the set of confounders, scale of the outcome variable and the form of the conditional expectations. For ease of presentation, we only consider TT_a; modifications for TT_b and TT_x are mostly evident.

The choice of scale is important as it is well known that generally, if the common trend (2) holds for one scale, it can hold for affine, but not for nonlinear transformations. This makes the choice of scale of the dependent variable essential, a problem that in practice is often ignored. Unless the researcher has a strong opinion about the scale, it should be chosen data-adaptively. The selection of covariates, as said, should be driven by reasons of total versus partial TT estimation, the reduction of noise, and Assumption 1. While the first is fully up to the researcher's interest, the second should be limited to a few obvious cases, as any inclusion of covariates has implications for interpretation and estimation; the third should be done data-adaptively.

Although we will propose a feasible, computationally inexpensive procedure, both selection problems are theoretically intertwined. Therefore, the data-adaptive selections should be based on the same objective function and be considered as a simultaneous, joint selection problem. Since the objective is to comply with Assumption 1, we propose

$$\frac{1}{n_{1\bullet}} \sum_{i:\, D_{i\bullet}=1} \left\{ m_{1t}(x_{i\bullet}) - m_{0t}(x_{i\bullet}) - m_{1(t-1)}(x_{i\bullet}) + m_{0(t-1)}(x_{i\bullet}) \right\}^2, \tag{9}$$

where the summation is over treated individuals in time period t (i.e., $n_{1\bullet} = n_{dt}$, $D_{i\bullet} = D_{it}$ and $x_{i\bullet} = x_{it}$) if we are interested in TT_a. Similarly, $n_{1\bullet} = n_{1t} + n_{1(t-1)}$, $D_{i\bullet} = D_i$ and $x_{i\bullet} = x_i$ if we are interested in TT_b. 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. The selector chooses the transformation and covariates that minimize (9). As we observe all non-treatment outcomes Y^0 only prior to the treatment, we propose to apply it to $t = 0$. Alternatively, one may likewise integrate (9) over the x_{i1} of the treated in $t = 1$.

3.1 Scale selection

Finding the transformation of Y for fulfilling (2) corresponds to finding the correct scale. Consequently, it must be a strictly monotone transformation that still provides a reasonable interpretation. You

[9] In our applicaiton, all of our covariates are discrete and hence a local-polynomial estimator is not only unnecessary, it is infeasible.

may think of the Box-Cox transformation which depends on a parameter, say θ. For each set x^S of covariates, there exists a parameter value θ^S_{opt}, that optimizes the common trend condition. Clearly, that criterion estimates the squared deviations from Assumption 1, and can thus be understood as a measure of variation. However, variations are scale dependent, irrespective of whether Assumption 1 is fulfilled in the population or not. Therefore, we propose to adapt the criterion by accounting for the different variances of $Y(\theta)$, and define for any given S, the optimal transformation parameter for Y by

$$\theta^S_{opt} = \underset{\theta}{argmin}\ \frac{1}{n_{1\bullet}} \sum_{i:\ D_{i\bullet}=1} \left\{ \widehat{m}_{1t}(x^S_{i\bullet}) - \widehat{m}_{0t}(x^S_{i\bullet}) - \widehat{m}_{1(t-1)}(x^S_{i\bullet}) + \widehat{m}_{0(t-1)}(x^S_{i\bullet}) \right\}^2 \Big/ \widehat{Var}_\bullet[Y(\theta)], \quad (10)$$

for $t = 0$ and where $\widehat{Var}_\bullet[Y(\theta)]$ is a standard estimator of the unconditional variance of the transformed responses when using θ. As for $n_{1\bullet}$ and $D_{i\bullet}$, the $\bullet$ indicates if this variance refers to the (sub-)population of all the treated or only the treated in $t = 0$.

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 a theoretically optimal bandwidth. This last statement, 'importantly change', 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 the context of such a selection procedure. 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 at least a very similar one) as the optimal bandwidth would. We suggest using the computationally attractive plug-in bandwidths as outlined in Henderson and Parmeter [2015] and/or Chu et al. [2015]. For small samples, these tend to slightly oversmooth what in addition would stabilize the numerical performance of the selection procedure. You should not search for the optimal bandwidth using a criterion like (9) or (10) as these criteria are supposed to be based on reasonable estimates of $\widehat{m}(\cdot)$, but not vice-versa.

3.2 Selection of covariates

While it is important to clarify which covariates to include, data-adaptive methods can be used for the confounders. 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. In the economics literature, it is often argued that the covariates should not be impacted themselves by the treatment, and therefore, only time invariant covariates are considered, or alternatively, only values of X observed before treatment.[10] This is actually not required, recall our assumptions above. So both, the set of covariates you definitely want to include, as well as the set of potential confounders you allow for, both depend on the parameter of interest. The correct interpretation hinges on your assumptions. These must be consistent with your data, and that your interpretation is consistent with these assumptions (see Kahn-Lang and Lang [2019]).

[10] In other fields people are often interested in direct or marginal effects and therefore include certain covariates because they are affected by treatment.

Suppose you have decided on your sets S of covariates from which you wish to select the set which best complies with Assumption 1. For the θ^S_{opt} defined above,

$$S_{opt} = \underset{S}{argmin} \frac{1}{n_{1\bullet}} \sum_{i:\, D_{i\bullet}=1} \left\{ \widehat{m}_{1t}(x^S_{i\bullet}) - \widehat{m}_{0t}(x^S_{i\bullet}) - \widehat{m}_{1(t-1)}(x^S_{i\bullet}) + \widehat{m}_{0(t-1)}(x^S_{i\bullet}) \right\}^2 / \widehat{Var}_\bullet[Y(\theta^S_{opt})] \quad (11)$$

defines the optimal set of covariates. As this is the analogue to (10), you (jointly) calculate the same criterion for all (θ, S) combinations to obtain $(\theta^{S_{opt}}_{opt}, S_{opt})$ which provides us with a setup for a DiD analysis with the most credible (of those proposed) identifiability assumptions. To be clear, you should not think of it as a step-wise elimination of covariates but rather of a comparison of all eligible sets of covariates.

If initially there are too many sets, one can perform pre-selection procedures. A simple method is a visual check to see to what extent a covariate is indeed a confounder. When plotting the distribution of a potential confounder per group and year, these should exhibit different features [Li, 1996], either between groups or else between years; otherwise they are not confounders. Pre-selection procedures can also be based on variable selection in regression; if they exhibit no impact on Y, they are of no use. In the context of nonparametric estimation, however, these pre-selection procedures are not less complex than directly applying our procedure [Hall et al., 2007]. Moreover, keep in mind that those are based on objective functions different from minimizing the deviation in (9). Generally we would advise against mixing different objective functions when the objective is actually the same as this has implications for correct interpretation.

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 (11), we propose to add the following penalty term

$$\left(2(k+p)^2 + 2(k+p)\right) / \left(n_{1\bullet} - (k+p)\right), \quad (12)$$

to penalize against adding too many covariates. As we will see in our simulations (Section 6), this factor helps correctly identify models with irrelevant covariates even with relatively small samples.

3.3 Testing a given combination

It is possible to formally conduct a nonparametric significance test to see if Assumption 1 is rejected for any given pair of (θ, S) for the period before treatment.[11] This can be done by taking (9) as the test statistic (for $t = 0$), and see if it differs significantly from zero. Its asymptotic properties and a bootstrap procedure to construct p-values are presented in Section 5 where we also apply the statistic for testing the significance of treatment effects. However, we can only test the credibility of Assumption 1, not the assumption itself. Extending this idea to post-treatment periods is questionable (cf., Kahn-Lang and Lang [2019]).

[11] In practice, you would test this at the "optimal set".

4 Treatment effect estimators

4.1 Conditional treatment effect on the treated

To simplify notation, Y and X now denote the adequately scaled response and the selected covariates. Recalling Section 2.2, and assuming the residuals of all groups are independent, estimators $\widehat{m}_{dt}(x)$, $d = 0, 1$, $t = 0, 1$ will be independent for any x from the common support as well. Then, the DiD estimators of conditional TT, namely

$$\widehat{TT}_x = \{\widehat{m}_{11}(x) - \widehat{m}_{01}(x)\} - \{\widehat{m}_{10}(x) - \widehat{m}_{00}(x)\}, \tag{13}$$

have (first-order) smoothing biases of order $O(h^2)$, which are the difference of differences of the corresponding individual biases given in (7) (i.e., $\{B_{11}(x) - B_{01}(x)\} - \{B_{10}(x) - B_{00}(x)\}$) where we implicitly allow each bias term to have its own set of bandwidths (h_{dt}, λ_{dt}). 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 (we assume they converge at the same rate).

Following (5), $\widehat{TT}_x$ converges at this rate to a normal distribution. As intuitively the direction and smoothness of the four $m_{dt}(\cdot)$ should not change over d and t, equation (7) suggests that the differencing has not only a bias reducing effect regarding identification (i.e., regarding a potential specification bias), but also regarding the smoothing bias. Unfortunately, the asymptotics of the more popular unconditional TT are not so straightforward. We detail these below.

In practice, no one would try to estimate the exact 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, in practice, we will estimate the variance via a wild bootstrap (outlined below).

Before we turn to the unconditional treatment effects, it is worth to recalling two points. First, looking at conditional treatment effects may be the most insightful way to study (potential) heterogeneity of treatment effects. Therefore we consider the above results not just as an intermediate step for the main result, and pay attention to TT_x in addition to TT_a and TT_b. Second, in the next subsection we directly integrate over all covariates x to obtain TT_a and TT_b. To further explore the heterogeneity of treatment effects, you may integrate over a subset 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 sex, you obtain $TT(x_2)$ to study the TT for men and women separately.

4.2 Unconditional treatment effect on the treated

Given the estimator in (13), it is straightforward to obtain a model-free DiD estimator for the unconditional TT by simply integrating $\widehat{TT}_x$. For sake of brevity, we consider TT_a, estimated by averaging over the (n_{11}) x_i observed in group $d = 1$ at time period $t = 1$:

$$\widehat{TT}_a = \frac{1}{n_{11}} \sum_{i:\, D_{i1}=1}^{n_{11}} \Big\{\widehat{m}_{11}(x_{i1}) - \widehat{m}_{01}(x_{i1}) - \widehat{m}_{10}(x_{i1}) + \widehat{m}_{00}(x_{i1})\Big\}. \tag{14}$$

This estimator is equal to $\widehat{TT}_b$ in a balanced panel where all covariates X_{it} are kept fixed over time. This does not imply that these variables are indeed time invariant, but that we only consider x-values observed at $t = 0$ (i.e., before treatment started).

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 Assumption 2 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, we continue with the above notation, assuming that in (14), we only average over interior points.

For sake of brevity, we do not derive in detail the asymptotics, but refer to the fact that in the case of independent residuals, statistic (14) 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 the kernel estimators $\widehat{m}_{dt}(x)$ in (5) imply we can choose λ and h for $dim(X^s) = p \leqslant 3$ such that $B = o(n_{dt}^{-1/2})$ and $\sqrt{n_{dt}h^p} = o(1)$. To achieve the same result for more than three continuous covariates, we can invoke higher-order kernels, or equivalently, local-polynomial estimators, both based on higher-order smoothness assumptions for $m_{dt}(\cdot)$ and the distributions of X.[12] Asymptotically, for $dim(X^c) = k$, we have no such restriction unless k increases with the sample size.

Formally, for $p \leqslant 3$ and with bandwidths properly chosen, the resulting first-order statistical properties are

$$\sqrt{n_{11}} \left\{ \frac{1}{n_{11}} \sum_{i:\, D_{i1}=1} \widehat{TT}_a(X_{i1}) - TT_a \right\} \to N(0, V_a), \tag{15}$$

where for $\kappa_{dt} = \lim(n_{dt}/n_{11})$, $\sigma^2_{dt}(x) = Var[Y|x, D = d, T = t]$, and $f_{dt}(x) = f(x|D = d, T = t)$. The variance is defined as

$$\begin{aligned} V_a \;=\; & E\left[\{m_{11}(X) - m_{10}(X) - m_{01}(X) + m_{00}(X) - TT\}^2 | D = T = 1\right] \\ & + E\left[\sigma^2_{11}(X)|D = T = 1\right] + E\left[\frac{\sigma^2_{10}(X) f^2_{11}(X)}{\kappa_{10} f^2_{10}(X)} | D = 1 - T = 1\right] \\ & + E\left[\frac{\sigma^2_{01}(X) f^2_{11}(X)}{\kappa_{01} f^2_{01}(X)} | D = 1 - T = 0\right] + E\left[\frac{\sigma^2_{00}(X) f^2_{11}(X)}{\kappa_{00} f^2_{00}(X)} | D = T = 0\right], \end{aligned} \tag{16}$$

where we emphasize the $\sqrt{n_{11}}$ rate.[13]

In Appendix A, we turn to the influence function (IF) to derive both (an alternative derivation to obtain) V_a and V_b (i.e., the analogous variance for $\widehat{TT}_b(X_{i1})$). If X does not change over time, the resulting simplified formula for $\widehat{TT}_b(X_{i1})$ coincides with the efficiency bounds derived in Sant'Anna and Zhao [2020], though in a quite different context (they introduce fully parametric doubly robust DiD estimation for time invariant X, where $D \perp T$, and $D \perp T|X$). Their study shows how our asymptotic results can be simplified when the data come from a balanced panel. For brevity, we show this simplification only for the more complex context of testing below.

[12] As we mentioned in the introduction, this is not a restrictive assumption for many economic data sets. The extension to larger numbers of continuous variables is still feasible, but requires additional assumptions.

[13] Uniform rates of convergence could be obtained by following results similar to Li and Racine [2007, pp. 78].

4.3 Bootstrap inference

In practice, asymptotic results for nonparametric statistics are rarely used directly for inference. Estimating any of the above variances is a nontrivial task that may involve several bandwidth choices, with the challenge that there hardly exist bandwidth selectors for such variance estimators. Even if you successfully estimate the above expressions, in finite samples, the suppressed remainder terms may still play a role, not to mention the slow convergence to normality.

In cases such as ours, the bootstrap is a widely accepted remedy. It is well known [Mammen, 1992], for nonparametric methods, 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 the kNN matching estimator. 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]. Further, Bodory et al. [2020] studied explicitly the consistency of the wild bootstrap for nonparametric matching estimators.

The distinction between the wild and nonparametric bootstrap methods often reduces to the question of how many moments are asymptotically matched. While asymptotic theory tells us that the higher these bootstrap residuals match the moments of the original residuals, the more efficient the bootstrap procedure, Davidson and Flachaire [2008] (see also the references therein) argue that you need quite large samples before these asymptotic results become effective. Following their recommendations, we propose a rather simple version (modifications towards higher-moment matching bootstraps are relatively straightforward). We first present the bootstrap procedure for continuous responses, and its analogue for discrete responses afterwards.

Given our consistent nonparametric estimators for (3), our residuals are given by

$$\widehat{u}_{it} = Y_{it} - \widehat{m}_{dt}(X_{it}) \ , \quad i = 1, ..., n_{dt}, \ d = 0, 1, \ t = 0, 1. \tag{17}$$

To obtain standard errors and confidence intervals for the TT estimators, we propose the following wild bootstrap. Generate $B \geqslant 100$ bootstrap samples $\{Y_{it}^{*b}, (D_{it} = d), t, X_{it}\}_{i=1}^{n_{dt}}$, $b = 1, ..., B$, for each of the four groups with $d = 0, 1$, $t = 0, 1$ by setting

$$Y_{it}^{*b} = \widehat{m}_{dt}(X_{it}) + u_{it}^{*b} \ , \text{ for a given } d, t, \ i = 1, ..., n_{dt}, \tag{18}$$

where u_{it}^{*b} can be generated by $\widehat{u}_{it}$ multiplied by an independent $N(0, 1)$ variable (which performed best in our simulations).[14] From these B tuples of the four samples, we calculate B estimators of $\widehat{TT}_x^{*b}$, $\widehat{TT}_a^{*b}$ and/or $\widehat{TT}_b^{*b}$, which are calculated as before (recall (13) and (14)), except that $\widehat{m}_{dt}(\cdot)$ is replaced with their bootstrap analogues $\widehat{m}_{dt}^{*b}(\cdot)$. From the B bootstrap estimates $\widehat{TT}_z^{*b}$ (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. When working with local-polynomial methods for estimating the conditional expectations, a link function, such as logit, is feasible for the binary case. A semiparametric bootstrap can be applied to draw from the conditional distribution defined by the choice of the link. More specifically, define a distribution $\mathcal{G}$ such that $Y|X = x \sim \mathcal{G}\{\eta(x)\}$; estimate the index

[14] Other authors favor the Radamacher distribution (in the context of significance tests in linear regression).

function $\eta(x)$ and its conditional expectation by local-likelihood, and draw the bootstrap responses Y_{it}^* from $\mathcal{G}\{\widehat{\eta}(X_{it})\}$ within each group (d,t).

If you use the local-constant version and face binary responses, as we do in our application, you can generate bootstrap replicates with randomly drawn $v^b \sim U[0,1]$ by

$$Y_{it}^{*b} := \mathbb{1}\{\widehat{m}_{dt}(X_{it}) > v^b\} \quad , \; b = 1, ..., B. \tag{19}$$

It is worth mentioning here that, in our application, we received essentially the same standard errors when applying bootstrap versions of (18) and (19).

5 Testing

To complete the cycle of our DiD analysis, we consider several testing problems of interest. We first briefly discuss how to test for significance of an unconditional treatment effect. We then introduce the aforementioned nonparametric test (Section 3.3) that can be used for supporting Assumption 1, as well as to jointly test for the significance of conditional treatment effects. This test is particularly interesting if treatment effect heterogeneity is large.

5.1 Significance of treatment effects

Let's first get the simplest cases out of the way. To test for significant treatment effects TT_z of type $z = x,\ a$ or b, we consider the null and alternative hypotheses

$$H_0^z : \; TT_z = 0 \qquad vs. \qquad H_1^z : \; TT_z \neq 0 \; . \tag{20}$$

Exploiting (15), along with what we know from Section 4.3, we can construct either an asymptotic or a bootstrap confidence interval to see if it includes zero. We reject the null if the confidence bound does not include zero. For these tests considered individually (for the case of TT_x, we are implying 'for a given x'), nothing remains to be shown, as the reader is already familiar with such procedures.

5.2 Composite significance testing in model-free DiD

As discussed in Section 3.3, while Assumption 1 cannot be directly tested, its credibility can. We can essentially use the same statistic as we did for the selection procedures, namely (9),[15] applied to the pre-treatment period (from $t = -1$ to 0), where by definition, $Y_i = Y_i^0$ for all subjects i. Formally, we suggest the general form test statistic

$$\mathcal{T}_t := \frac{1}{n_{1t}} \sum_{i:D_{1t}=1}^{n_{1t}} \left\{ \widehat{m}_{1t}(x_{it}) - \widehat{m}_{0t}(x_{it}) - \widehat{m}_{1(t-1)}(x_{it}) + \widehat{m}_{0(t-1)}(x_{it}) \right\}^2 , \tag{21}$$

[15] A rescaling by the response variance estimate is not needed as this was only done to make the statistics comparable over different transformations of Y.

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

$$H_0^t: \ M_t(x) := m_{1t}(x) - m_{0t}(x) - m_{1(t-1)}(x) + m_{0(t-1)}(x) = 0 \ \forall x \in supp(X|D=1, T=0) \ . \tag{22}$$

More formally, these include:

Bias stability condition To test the bias stability ('parallel trend') condition, we look at data prior to treatment (from period -1 to 0). Here we are interested in $\mathcal{T}_0$. This test statistic,

$$\mathcal{T}_0 := \frac{1}{n_{10}} \sum_{i:D_{10}=1}^{n_{10}} \left\{\widehat{m}_{10}(x_{i0}) - \widehat{m}_{00}(x_{i0}) - \widehat{m}_{1(-1)}(x_{i0}) + \widehat{m}_{0(-1)}(x_{i0})\right\}^2, \tag{23}$$

checks the credibility of (2) by considering the null hypotheses

$$H_0^0: \ M_0(x) := m_{10}(x) - m_{00}(x) - m_{1(-1)}(x) + m_{0(-1)}(x) = 0 \ \forall x \in supp(X|D=1, T=0) \ . \tag{24}$$

Joint significance of heterogenous effects When heterogeneity in treatment effects is important, it is much 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 that case,

$$\mathcal{T}_1 := \frac{1}{n_{11}} \sum_{i:D_{11}=1}^{n_{11}} \left\{\widehat{m}_{11}(x_{i1}) - \widehat{m}_{01}(x_{i1}) - \widehat{m}_{10}(x_{i1}) + \widehat{m}_{00}(x_{i1})\right\}^2, \tag{25}$$

is the appropriate test statistic for the null hypothesis (i.e., $M_1(x)$).[16]

Homogenous treatment effects You can extend the null hypotheses for testing the null H_0^z to $TT_z = c$ for $z = a, b, x$, with c being a given constant. The interesting case is when you apply this extension to test all TT_x jointly over the sample of interest. The resulting test statistic here is

$$\mathcal{T}_H := \frac{1}{n_{11}} \sum_{i:D_{1t}=1}^{n_{11}} \left\{\widehat{m}_{11}(x_{i1}) - \widehat{m}_{01}(x_{i1}) - \widehat{m}_{10}(x_{i1}) + \widehat{m}_{00}(x_{i1}) - c\right\}^2 \ . \tag{26}$$

For example, this can be employed to test for significant heterogeneity of treatment effects over X by setting $c := \widehat{TT}_a$. If $dim(x) = 1$, we could alternatively construct bootstrap confidence intervals and bands around TT_x for all x. In general (for our testing procedures), our bootstrap will require us to resample the data under the null and a simple bootstrap procedure such as that for TT_x is not sufficient.

5.3 Asymptotic behavior

In what follows, we study the asymptotic behavior of $\mathcal{T}_1$. For $\mathcal{T}_0$ and $\mathcal{T}_H$, the derivations follow analogously, noting that $\widehat{TT}_a$ converges faster than $\widehat{m}_{dt}(\cdot)$ such that its randomness is negligible in (26). 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 behavior of the test statistic with a balanced panel.

Recall our definition of κ_{dt} directly following (15). Lets define the four single dimension design densities $f_{dt}(x)$ implicitly by $\int_0^{x_{it}} f_{dt}(x)dx = i/n_{dt}$ for all observed x_{it} with $D_{it} = d$.[17] We assume all $m_{dt}(\cdot)$ and

[16] As you may prefer TT_b over TT_a, you can also average in (25) over all treated ($n_{11} + n_{10}$).

[17] We could have assumed all samples are asymptotically regular designs with respect to their density $f_{dt}(\cdot)$.

$f_{dt}(\cdot)$ are $r \geqslant 2$ times continuously differentiable on $[0,1]$. Accordingly, our kernel $W(X,x,h)$ is of order r. Consider the optimal testing rate $h = O(n_{11}^{-2/(4r+1)})$ with $n_{11}h^2 \to \infty$ (see Ingster [1993]). We obtain the result that

$$n_{11}\sqrt{h}\left\{\mathcal{T}_1 - \frac{1}{n_{11}h}\int W^2 \sum_{d,t=0}^{1}\int \frac{\sigma_{dt}^2(x)f_{11}^2(x)}{\kappa_{dt}f_{dt}(x)}dx\right\} \longrightarrow N(0,\mathcal{V}) \quad \text{as all } n_{dt}\to\infty, \tag{27}$$

where the variance $\mathcal{V}/(n_{11}^2h)$ of our statistic $\mathcal{T}_1$ is

$$\frac{2}{n_{11}^2h}\int (W*W)^2\left(\sum_{d,t=0}^{1}\int \frac{\sigma_{dt}^4(x)f_{11}^2(x)}{\kappa_{dt}^2f_{dt}^2(x)}dx + 2\sum_{mix(dt,ks)}\int \frac{\sigma_{dt}^2(x)\sigma_{ks}^2(x)f_{11}^2(x)}{\kappa_{dt}\kappa_{ks}f_{dt}(x)f_{ks}(x)}dx\right), \tag{28}$$

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

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 (27), (28), and in the definition of the κ_{dt}. Its extension to allow for the inclusion of weights and trimming is relatively straightforward, see Section 8.

The same calculations can be done 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 (27) 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, as in estimation, their inclusion does not change the asymptotic rate.

With balanced panels ($n_1 = n_{11} = n_{10}$ and $n_0 = n_{01} = n_{00}$), when considering $\mathcal{T}_1$, assuming $u_{i0} \perp u_{i1}$ for all i becomes less credible.[18] In this situation, the asymptotics of our test simplify even when $u_{i0} \not\perp u_{i1}$, as

$$\widehat{m}_{d1}(x) - \widehat{m}_{d0}(x) = \frac{\sum_{D_i=d:i=1}^{n_d} W_h(X_{i0}-x)\ (Y_{i1}-Y_{i0})}{\sum_{D_i=d:i=1}^{n_d} W_h(X_{i0}-x)}.$$

Consequently, with $f_1(\cdot)$ being defined as the density of X for the treated, $f_0(\cdot)$ for the control,

$$n_1\sqrt{h}\left\{\mathcal{T}_1 - \frac{\int W^2}{h}\int \frac{\tilde{\sigma}_1^2(x)f_1(x)}{n_1} + \frac{\tilde{\sigma}_0^2(x)f_1^2(x)}{n_0f_0(x)}dx\right\} \longrightarrow N(0,\tilde{\mathcal{V}}), \tag{29}$$

where $\tilde{\sigma}_d^2(x) = Var(u_{i1}-u_{i0}|X_{i0}=x)$, and for $\kappa_d = \lim(n_d/n_1)$

$$\tilde{\mathcal{V}} = 2\int (W*W)^2\left(\sum_{d=0}^{1}\int \frac{\tilde{\sigma}_d^4(x)f_1^2(x)}{\kappa_d^2f_d^2(x)}dx + 2\int \frac{\tilde{\sigma}_1^2(x)\tilde{\sigma}_0^2(x)f_1(x)}{\kappa_1\kappa_0f_0(x)}dx\right). \tag{30}$$

5.4 Feasible bootstrap tests

Arguments in favor of using a bootstrap for testing are even stronger (because $\mathcal{T}_t$ is a nonparametric test). We need large samples before the first-order terms fully dominate the second and third-order terms.[19] Even if the samples were large enough to trust the normal approximation and we could neglect

[18] In the case of repeated cross-sections, we typically observe u_{i0} and u_{j1}, where $i \neq j$, in general. In other words, dependencies in residuals over time are excluded as we have cohorts.

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

higher-order terms, estimation of the first-order terms would still remain a non-trivial problem. We propose to approximate the critical value of the test via a bootstrap.

The challenge is to simulate the distribution of the statistic, say $\mathcal{T}_1$, 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 $H_0^1: \ M_1(x) = 0$ for all x of interest, see (24). Other proposals may be feasible, but ours follows ideas we found in the most related literature on nonparametric testing, namely Dette and Neumeyer [2001] and Vilar and Vilar [2012]. The latter provides 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 procedure. In our context, different scenarios are conceivable to comply with H_0^1. For that reason, we need to take the residuals for the bootstrap DGP from the alternative (as proposed by Vilar and Vilar [2012]) instead of from the null model (as proposed by Dette and Neumeyer [2001]). This has consequences for the calibration [Sperlich, 2014]; see below. The four step bootstrap procedure proceeds as follows:

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 $B \geqslant 100$ 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

$$Y_{it}^{*b} = \widehat{m}_t(X_{it}) + u_{it}^{*b} \ , \text{ for given } d,t, \ i=1,...,n_{dt}, \tag{31}$$

   where u_{it}^{*b} might be generated by $\widehat{u}_{it}$ from (17) times an independent $N(0,1)$ variable, see discussion in Section 4.3.

3. From these B tuples of four samples, calculate B estimators $\mathcal{T}_1^{*b}$ which are calculated as in (23), 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 is the pooling in step 1, which guarantees that the null hypothesis (2) 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, unfortunately, has the tendency to suffer from size distortions in the sense of over-rejection. The reason why our proposal generally outperforms the latter is the following: D is definitely a function of X (by the definition of confounders), T should not be. 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 does not always have to be like this, but it is much more likely than not. This was confirmed by many simulations (see Section 6). For huge data sets, we may work with the asymptotic expressions and estimate the bias and variance given in (27) and (28) or the much simpler versions (29) and (30) in case of balanced panels.

It is obvious how to modify this procedure for $\mathcal{T}_0$, but we must be careful; consistency of this bootstrap test does not necessarily carry over to all kind of modifications or generalizations. Neumeyer and Sperlich [2006] studied a similar test, comparing marginal, additive separable impacts. In their paper, this bootstrap procedure was not only inconsistent, but diverged.

For nonparametric analysis of continuous covariates, Faraway [1990] and Härdle and Marron [1991] notice that those bootstrap procedures do not consistently capture the smoothing bias. They propose 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, bias reduction or a double bootstrap. Neumann and Polzehl [1998] show that asymptotically, using local-polynomials with undersmoothing h works as well, as the bias converges faster.

6 Simulations

In this section, we 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. 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 $\epsilon_{it} \sim N(0, 1.5)$, and $u_{it} \sim N(0, \sigma_u^2)$ for $t = -1, 0, 1$. We obtain the treatment status and outcome values as:

$$D_{it} = 1\{0.75X_{it,1} - 0.5X_{it,2}^2 > \epsilon_{it}\} \tag{32}$$

$$Y_{it} = 1 + t(2 + X_{it,1} + X_{it,2}^2) + D_{it} + D_{it}1\{t \geqslant 1\} + u_{it} \tag{33}$$

where the treatment effect on the treated is the coefficient on the interaction term (i.e., $TT = 1.0$) in (33).[20] In (33) 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.

We emphasize here, while we choose $n = 100, 200, 400$ and 800, the effective sample sizes are much smaller. The last two columns of numbers in Table 1 give the average sample size (to the nearest integer) for n_{10} (the number of observations we sum over in (9)), and the smallest sample size over all n_{dt} ($d \in \{0, 1\}, t \in \{-1, 0, 1\}$).[21] 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 we will see that our methods still perform admirably.

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

[21] The effective sample sizes are nearly identical in the remaining tables of this section.

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 in this setting (i.e., stickman comparison models). Further, should the parametric models be correctly specified, 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.[22]

6.1 Choice of the confounder set

To see if our method appropriately picks the correct set of covariates, we generate our data as in (33). 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, each of 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 $j = 1, 2, \ldots, 5$. More formally, we consider the following sets: $S_{1,2} = \{X_{it,1}, X_{it,2}\}$, $S_{1,3} = \{X_{it,1}, X_{it,3}\}$, $S_{2,4} = \{X_{it,2}, X_{it,4}\}$, $S_{3,4} = \{X_{it,3}, X_{it,4}\}$, $S_{4,5} = \{X_{it,4}, X_{it,5}\}$, $S_{1,3,4} = \{X_{it,1}, X_{it,3}, X_{it,4}\}$, $S_{2,4,5} = \{X_{it,2}, X_{it,4}, X_{it,5}\}$, $S_{1,2,3} = \{X_{it,1}, X_{it,2}, X_{it,3}\}$, $S_{1,2,4} = \{X_{it,1}, X_{it,2}, X_{it,4}\}$, $S_{1,2,3,4} = \{X_{it,1}, X_{it,2}, X_{it,3}, X_{it,4}\}$, and $S_{1,2,3,4,5} = \{X_{it,1}, X_{it,2}, X_{it,3}, X_{it,4}, X_{it,5}\}$. 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 (12) 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 the fraction correct should be 1. Table 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.

As expected, the first column of numbers represent the hardest case. With $n = 100$ (i.e., some n_{dt} only above 10), we are roughly at or above random choice. For $n > 100$, it improves even for low signal-to-noise ratios.[23] 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

[22] We will compare our methods to linear parametric methods in our empirical application.

[23] We continued to raise the sample size to ensure that these fractions tended towards 1.000. When doubling the sample size, this occurred by $n = 3200$ (approximately $n_{10} = 575$) for the case where $\sigma_u^2 = 2.0$.

Table 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 ($n = \Sigma_{t=-1}^{1} n_t = \Sigma_{d=0}^{1}\Sigma_{t=-1}^{1} n_{dt}$)

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

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.

6.2 Test

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 4.3). We conduct our simulations along the problem of studying the bias stability ('parallel path') condition.[24] We generate our data as in (33) to determine the size of the test. To determine the power, we change the indicator function to $1(t \geqslant 0)$ in (33) 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 Section 5.4, 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 as shown in Section

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

5.2 ($\mathcal{T}_0$ in (23)), but calculate the residuals using the bandwidth procedure of Vilar and Vilar [2012].[25] 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).

The results for both the size and power of our test ($\mathcal{T}_0$ in 23) can be found in Table 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 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.

Table 2: Size and power of our bias stability ('parallel path') condition test ($\mathcal{T}_0$ in (23)): 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				
	1%	5%	10%	1%	5%	10%	n_{10}	$\min(n_{dt})$
$\sigma_u^2 = 1.0$								
$n = 100$	0.006	0.036	0.070	0.067	0.212	0.326	18	13
$n = 200$	0.009	0.055	0.086	0.190	0.401	0.531	35	28
$n = 400$	0.011	0.054	0.121	0.488	0.743	0.837	71	58
$n = 800$	0.010	0.049	0.109	0.870	0.971	0.987	140	119
$\sigma_u^2 = 2.0$								
$n = 100$	0.006	0.026	0.083	0.030	0.127	0.213	18	12
$n = 200$	0.012	0.042	0.128	0.074	0.222	0.332	35	27
$n = 400$	0.009	0.056	0.125	0.212	0.420	0.573	71	58
$n = 800$	0.011	0.053	0.110	0.517	0.724	0.823	140	119

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.

[25] 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 prefer the automated approach in Vilar and Vilar [2012]. These results are available upon request.

6.3 Performance of the treatment effect estimator

Finally, we move to estimates of the TT itself as well as 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).[26] 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 4.3. We do not require a bandwidth g and use h for both estimation and in our bootstrap.[27]

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

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

Table 3 gives the results from our simulations. We again choose a moderate (upper panel) and a low (lower panel) signal-to-noise ratio. In each case, the finite sample bias exists and tends towards zero as n increases. Again, larger biases are a function of using plug-in bandwidths which tend to oversmooth (LSCV bandwidths leads to much smaller average biases).[28] 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

[26] 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.

[27] 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.

[28] We advocate for using cross-validated bandwidths in practice when estimating the TT. The sign of the bias is not random, but why it is negative can only be deduced from the average over the linear combinations of individual biases $B_{dt}(x, \lambda, h)$, see (7), as discussed after definition (13), which in turn depends on the particular bandwidth choices, true densities and functions. Importantly, it is minor in size and rapidly converges to zero.

the bias). The third column of numbers give 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.

7 Human capital responses to the Deferred Action for Childhood Arrivals program

On June 18th, 2020, the Supreme Court of the United States ruled that the president could not immediately end the Deferred Action for Childhood Arrivals (DACA) program. 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. Presumably 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. These results are promising, but they rely on restrictive assumptions and hence are subject to misspecification bias and potential inconsistency.

7.1 Data

The data come directly from Kuka et al. [2020] and we will only discuss them briefly.[29] 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.[30] 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[31] and post-secondary attendance (three different binary left-hand-side variables).[32]

The ACS includes a large amount of demographic variables which are exploited by Kuka et al. [2020] to attempt to make Assumption 1 hold. Specifically, they account for fixed individual characteristics by including controls for sex, year of immigration and birth region. Given the nature of the parametric models, they also include interactive dummies for age of immigation-by-eligibility and age-by-eligibility

[29] The data are freely available online via https://doi.org/10.1257/pol.20180352 .

[30] They also look at immigrants aged 23–30 as a type of falsification test.

[31] High school completion includes both those who graduated from high school as well as those who earned a passing grade on the General Educational Development (GED) test.

[32] In the presence of a binary outcome variable, linear parametric difference-in-differences estimators do not guarantee that the predicted expected potential outcomes respect this support condition. Our nonparametric estimator, in contrast, will guarantee this support condition.

fixed effects. Further, 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 does include these as special cases. 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.[33,34]

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. Further, the Census Bureau also makes the survey available in Spanish.

It is also important to point out that Kuka et al. [2020] note that their measure of eligibility is measured with noise as it includes noncitizens who may have green cards or may be temporary visa holders and hence not eligibile for DACA. Hence, 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 of DACA by roughly 45 percent.

7.2 Empirical results

The parametric results can be found in Tables 4 and 5. These correspond to their model

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

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 one 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 4), we use cross-validated bandwidths (Section 8.1) and use our bootstrap procedure (with $B = 999$ bootstrap samples) to calculate our standard errors (Section 4.3).

The final three values associated with each sample in Tables 4 and 5 are the sample size, the mean of the outcome variable and the p-value associated with our bias stability ('parallel path') condition test.

[33] In the Hispanic sample and in the high-take up sample, we exclude the variable for race.

[34] If we know the individuals age, the year they immigrated and the age they were when they immigrated, we know the current year. Kuka et al. [2020] used interactions to keep all three. We estimated our models without the year variable. Out of curiosity, we also tried calculating the bandwidths for the models including the year variable. The cross-validation procedure correctly smoothed out the year variable [Hall et al., 2007].

The bias stability condition test shows mixed results.[35] In Table 4, we firmly reject the null that the 'parallel path' is present in our sample for 14-18 year olds, but are unable to reject it for each case for 19-22 year olds. Table 5 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.[36] As such, we 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 attempt to satisfy the identification condition.

7.2.1 School attendance

The results for school attendance are found in Table 4. 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 (albiet 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.

Table 4 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 the impacts of DACA on this group. The parametric model gives positive, but insigificant estimates. On the other hand, 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. For example, Hsin and Ortega [2018] found that DACA increased dropout rates by 7.3% in 2018. Similarly, Amuedo-Dorantes and Antman [2017] found that DACA reduced the probability of school enrollment of eligible higher-educated individuals as it incresed the likelihood of employment of men, in particular. The explanation there was that the lack of authorization led inviduals 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.

7.2.2 High school completion and college enrollment

The effects of DACA on high school completion and college enrollment can be found in Table 5. The first three columns of represent the effect of DACA on high school completion (GED or diploma) for all immigrants, Hispanic immigrants and immigrants from high take-up countries, respectively. These

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

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

Table 4: Effect of DACA on school attendance

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

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 (19, 19-22 and 23-30).

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. Taking these results at face value suggests that the impact appears to be stronger for younger individuals.

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. Although these differences are relatively large, 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.

Table 5: Effect of DACA on high school completion and college enrollment

	High-School			College		
	All	Hispanic	High take-up	All	Hispanic	High take-up
Age 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
Age 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
Age 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

A similar patter occurs for the impact of DACA on some college. The fourth through sixth columns of Table 5 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, before we read too much into these results, it is important to note that 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 model was 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.[37] 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).

8 Implementation

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

8.1 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.

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.

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.[38] 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.,

$$LSCV(h, \lambda) = \sum_{i:D_{i1}=1}^{n_{11}} \left(Y_i - \widehat{E}_{-i}\left[Y_i | X = x_i\right] \right)^2, \tag{34}$$

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 will share the same smoothness as the first).

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

[38] We tried this in 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 (34) 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 8.3, 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 the plug-in bandwidths to calculate the relevant test statistics.

In the case of testing with continuous covariates, as we discussed in Section 5.4, we suggest an approach analogous to that in Vilar and Vilar [2012], whereby they search for the bandwidth over the set of covariates X that is the largest (h), and use that bandwidth to smooth the remaining covariates (g). This is simple, but works well in simulations and is preferable to the common practice of multiplying h by a fixed constant (e.g., $g = 1.5h$). Note that as we only have discrete data in our application (Section 7), we do not face the choice of h versus g in our test statistics.

8.2 Sampling weights

In our application, sample weights are used. This can be implemented in the most generic setting of (6). 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 $\sum_{i=1}^{n} Y_i w_i K(X_i, x, h, \lambda)$ and/or $\sum_{i=1}^{n} w_i K(X_i, x, h, \lambda)$ 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).

8.3 Algorithm and coding

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, estimation, and testing, as it may be desirable to disentangle them in an application. 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 in (2). 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.[39]

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 the conditional expectations for the other three cases ($d = 1, t = -1$, $d = 0, t = 0$ and $d = 0, t = -1$).[40]

4. For each combination listed in the previous step, calculate the bias stability condition in (2). 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 $d = 0, t = 0$).[41]

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 4.3 and perform the tests of interest from Section 5.2.

The first two procedures require data prior to period 0 whereas the third does not. The first procedure `bsc.choice()`, identifies the set of covariates and scale of the outcome variable that minimize the objective functions in (10) and (11) jointly. The procedure `bsc.test()`, tests if the bias stability condition is violated. The procedure `npdid.estimation()`, estimates the treatment effects. `bsc.test()` can be used again to test for significant treatment effects. All procedure code is described in greater detail in Appendix C.

8.4 Parametric and semi-parametric alternatives

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

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

[40] 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). This step requires that past information is available, notably at least $t = -1$.

[41] For continuous variables, $h_j = c_j \hat{\sigma}_{x_j} n^{-1/(4+q)}$, where c_j is the scale factor and $\hat{\sigma}_{x_j}$ 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.

that our method be first. Our methods do not have to be the last step, instead, they can guide the practitioner to find appropriate models 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.[42]

9 Conclusions and directions for further extensions

We suggest a general framework for causal analysis (with covariates) via model-free DiD estimation and testing. We showed 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 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 were verified by simulation studies under rather complex designs.

We applied our techniques to study the impact of DACA on human capital decisions. We compared 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 found a positive (albeit larger) impact of DACA on high school attendance and high school completion, but we found that they were unable to identify the negative impact of DACA on school enrollment of college aged individuals.

One may ask about post-selection inference as we propose a procedure that allows us to select between different sets of covariates and scales of Y. This is not directly comparable with the existing literature on variable selection or so-called double machine learning. Our criterion is not the covariates contribution to a regression, but the maximization of bias stability, and therefore falls rather in the class of so-called *selected inference problems*, which is more intricate and perhaps a field for future research. It tries to estimate standard errors and p-values, conditioned on your choices. Alternatively, in the spirit of an explorative nonparametric analysis, one may consider it a problem of the post-selection inference (*PoSI*) that suggests to account for all variation of the entire statistical analysis. There we estimate the variance of a final estimator or test statistic allowing for any choice weighted with its likelihood. We do not condition on choices (typically estimated via Monte Carlo). In our case, you could construct an outer bootstrap loop that runs over all steps. These procedures are not very popular in practice as they are costly and can give unreasonably large standard errors.

[42] 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 $p \leqslant 3$ restriction.

References

A. Abadie. Semiparametric Difference-in-Differences Estimators. *Review of Economic Studies*, 72(1): 1–19, 01 2005.

A. Abadie and G. W. Imbens. On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6):1537–1557, 2008.

C. Amuedo-Dorantes and F. Antman. Schooling and labor market effects of temporary authorization: Evidence from daca. *Journal of Population Economics*, 30:339–373, 2017.

D. Ang. 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, July 2019.

H. Bodory, L. Camponovo, M. Huber, and M. Lechner. The finite sample performance of inference methods for propensity score matching and weighting estimators. *Journal of Business & Economic Statistics*, 38(1):183–200, 2020.

R. Cao-Abad and W. González-Manteiga. Bootstrap methods in regression smoothing. *Journal of Nonparametric Statistics*, 2(4):379–388, 1993.

C.-Y. Chu, D. J. Henderson, and C. F. Parmeter. Plug-in bandwidth selection for kernel density estimation with discrete data. *Econometrics*, 3(2):199–214, 2015.

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

H. Dette and N. Neumeyer. Nonparametric analysis of covariance. *Annals of Statistics*, 29(5):1361–1400, 2001.

J. J. Faraway. Bootstrap selection of bandwidth and confidence bands for nonparametric regression. *Journal of Statistical Computation and Simulation*, 37(1-2):37–44, 1990.

M. Frölich. Matching estimators and optimal bandwidth choice. *Statistics and Computing*, 156:197–215, 2005.

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

P. Hall, Q. Li, and J. S. Racine. Nonparametric estimation of regression functions in the presence of irrelevant regressors. *Review of Economics and Statistics*, 89(4):784–789, 2007.

W. Härdle and J. S. Marron. Bootstrap simultaneous error bars for nonparametric regression. *Annals of Statistics*, 19(2):778–796, 1991.

W. Härdle and T. M. Stoker. Investigating smooth multiple regression by the method of average derivatives. *Journal of the American Statistical Association*, 84(408):986–995, 1989.

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

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

A. Hsin and F. Ortega. The effects of deferred action for childhood arrivals on the educational outcomes of undocumented students. *Demography*, 55:1487–1506, 2018.

S. Jayachandran, A. Lleras-Muney, and K. V. Smith. 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, April 2010.

A. Kahn-Lang and K. Lang. 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, 2019.

E. Kuka, N. Shenhav, and K. Shih. Do human capital decisions respond to the returns to education? evidence from DACA. *American Economic Journal: Economic Policy*, 12(1):293–324, February 2020.

M. Lechner. The estimation of causal effects by difference-in-difference methods. *Foundations and Trends® in Econometrics*, 4(3):165–224, 2011.

Q. Li. Nonparametric testing of closeness between two unknown distribution functions. *Econometric Reviews*, 15(3):261–274, 1996.

Q. Li and J. S. Racine. *Nonparametric Econometrics: Theory and Practice*. Princeton University Press, 2007.

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

D. McKenzie, C. Theoharides, and D. Yang. 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, April 2014.

B. D. Meyer. Natural and quasi-experiments in economics. *Journal of Business & Economic Statistics*, 13(2):151–161, 1995.

M. H. Neumann and J. Polzehl. Simultaneous bootstrap confidence bands in nonparametric regression. *Journal of Nonparametric Statistics*, 9(4):307–333, 1998.

N. Neumeyer and S. Sperlich. Comparison of separable components in different samples. *Scandinavian Journal of Statistics*, 33:477–501, 2006.

D. Ouyang, Q. Li, and J. S. Racine. Nonparametric estimation of regression functions with discrete regressors. *Econometric Theory*, 25(1):1–42, 2009.

M. Panhans. Adverse selection in aca exchange markets: Evidence from colorado. *American Economic Journal: Applied Economics*, 11(2):1–36, April 2019.

C. Parmeter, Z. Zheng, and P. McCann. Cross-validated bandwidths and significance testing. *Advances in Econometrics*, 25:71–98, 2009.

D. N. Politis. Model-free model-fitting and predictive distributions. *TEST*, 22:183–221, 2013.

N. G. Pope. The effects of dacamentation: The impact of deferred action for childhood arrivals on unauthorized immigrants. *Journal of Public Economics*, 143:98–114, 2016.

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

J. Roca-Pardiñas and S. Sperlich. Feasible estimation in generalized structured models. *Statistics and Computing*, 20:367–379, 2010.

S. Ruggles, K. Genadek, R. Goeken, J. Grover, and M. Sobek. Integrated public use microdata series: Version 7.0 [dataset], 2018.

P. H. Sant'Anna and J. Zhao. Doubly robust difference-in-differences estimators. *Journal of Econometrics*, 219(1):101–122, 2020.

S. J. Sheather and M. C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 53(3):683–690, 1991.

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

T. Słoczyński. A general weighted average representation of the ordinary and two-stage least squares estimands. IZA Discussion Paper No. 11866, 2018.

S. Sperlich. Comments on: Model-free model-fitting and predictive distributions. *TEST*, 22:227–233, 2013.

S. Sperlich. On the choice of regularization parameters in specification testing: A critical discussion. *Empirical Economics*, 47:427–450, 2014.

J. M. Vilar and J. A. Vilar. A bootstrap test for the equality of nonparametric regression curves under dependence. *Communications in Statistics - Theory and Methods*, 41(6):1069–1088, 2012.

J. M. Vilar-Fernández and W. González-Manteiga. Nonparametric comparison of curves with dependent errors. *Statistics*, 38(2):81–99, 2004.

A Influence functions

We may look at the influence function (TT_a) which for $p_{dt}(x) = Pr(D = d, T = t|x)$ can be written as

$$\begin{aligned}\varphi_a(X) = \tfrac{DT}{E[DT]}\left[m_{11}(X) - m_{10}(X) - \{m_{01}(X) - m_{00}(X)\} - TT_a\right] \\ + \tfrac{DT}{E[DT]}\{Y - m_{11}(X)\} - \tfrac{D(1-T)}{E[DT]}\tfrac{p_{11}(X)}{p_{10}(X)}\{Y - m_{10}(X)\} \\ - \tfrac{(1-D)T}{E[DT]}\tfrac{p_{11}(X)}{p_{01}(X)}\{Y - m_{01}(X)\} + \tfrac{(1-D)(1-T)}{E[DT]}\tfrac{p_{11}(X)}{p_{00}(X)}\{Y - m_{00}(X)\} + R_{h,n_{11}}(X),\end{aligned} \tag{A1}$$

where $R_{h,n_{11}}(X)$ is a remainder term due to the nonparametric estimates $\widehat{m}_{dt}(\cdot)$. Note that we used $E[D(1-T)p_{11}(X)p_{10}^{-1}(X)] = E[(1-D)Tp_{11}(X)p_{01}^{-1}(X)] = E[(1-D)(1-T)p_{11}(X)p_{00}^{-1}(X)] = E[DT]$. Noting that $n_{11} = n\ E[DT]$, we immediately get the seemingly simpler variance representation (cf., 16)

$$\begin{aligned}V_a \;=\; & \frac{1}{E[DT]}E\Bigg[p_{11}(X)\{m_{11}(X) - m_{10}(X) - m_{01}(X) + m_{00}(X) - TT\}^2 \\ & + p_{11}(X)\sigma_{11}^2(X) + \frac{p_{11}^2(X)}{p_{10}(X)}\sigma_{10}^2(X) + \frac{p_{11}^2(X)}{p_{01}(X)}\sigma_{01}^2(X) + \frac{p_{11}^2(X)}{p_{00}(X)}\sigma_{00}^2(X)\Bigg]\ .\end{aligned} \tag{A2}$$

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

$$\begin{aligned}\varphi_b(X) = \tfrac{D}{E[D]}\left[m_{11}(X) - m_{10}(X) - \{m_{01}(X) - m_{00}(X)\} - TT\right] \\ + \tfrac{DT}{E[DT]}\{Y - m_{11}(X)\} - \tfrac{D(1-T)}{E[D(1-T)]}\{Y - m_{10}(X)\} \\ - \tfrac{(1-D)T}{E[DT]}\tfrac{p(X)}{1-p(X)}\{Y - m_{01}(X)\} + \tfrac{(1-D)(1-T)}{E[D(1-T)]}\tfrac{p(X)}{1-p(X)}\{Y - m_{00}(X)\} + R_{h,n_1}(X)\quad .\end{aligned} \tag{A3}$$

Consequently, in (15), $n_1 = n_{11} + n_{10}$ replaces n_{11} and the variance becomes

$$\begin{aligned}V_b \;=\; & Var(\widehat{TT}_b) = \frac{1}{n}E\Bigg[\frac{p(X)}{E^2[D]}\{m_{11}(X) - m_{10}(X) - m_{01}(X) + m_{00}(X) - TT\}^2 \\ & + \frac{p_{11}(X)}{E^2[DT]}\sigma_{11}^2(X) + \frac{p_{10}(X)}{E^2[D(1-T)]}\sigma_{10}^2(X) \\ & + \frac{p_{01}(X)}{E^2[DT]}\frac{p^2(X)}{\{1-p(X)\}^2}\sigma_{01}^2(X) + \frac{p_{00}(X)}{E^2[D(1-T)]}\frac{p^2(X)}{\{1-p(X)\}^2}\sigma_{00}^2(X)\Bigg]\end{aligned} \tag{A4}$$

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 $(n_{11} + n_{10})^{-1}$. It should be clear that (A4) simplifies if $D \perp T$ or/and $D \perp T|X$. Furthermore, 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 (A4), note that $p_{1t}(x) = p(x)\ Pr(T = t|D = 1, x)$ and $p_{0t}(x) = \{1 - p(x)\}Pr(T = t|D = 0, x)$. The resulting simplified formula of (A4) coincides with the efficiency bounds derived in Sant'Anna and Zhao [2020].

B Technical proof for the test statistic

Here we briefly indicate the main ideas 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

$$\mathcal{T}_1 = \sum_{d,t=0}^{1} \Gamma_{dt} + 2 \sum_{mix(dt,ks)} (-1)^{d+k+t+s} \Gamma_{dt,ks} + o_P\left(\frac{1}{n_{11}\sqrt{h}}\right), \tag{B1}$$

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

$$\Gamma_{dt} = \sum_{D_i=d:i=1}^{n_{dt}} \sum_{D_j=d:j=1}^{n_{dt}} \int W_{dt}(x_{it}) W_{dt}(x_{jt}) dF_{11}(x) \ u_{it}u_{jt} \tag{B2}$$

$$\Gamma_{dt,ks} = \sum_{D_i=d:i=1}^{n_{dt}} \sum_{D_j=k:j=1}^{n_{ks}} \int W_{dt}(x_{it}) W_{ks}(x_{js}) dF_{11}(x) \ u_{it}u_{js}, \tag{B3}$$

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 reconsider below for balanced panels, 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 (B1), but suppressing $\sum_{D_i=d:i=1}^{n_{dt}} \int W_{dt}^2(x_{it}) dF_{11}(x) \ u_{it}^2$ in the Γ_{dt}. 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 (28) 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 (28) containing all mixtures $mix(dt, ks)$ is

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

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].

C Procedure code

In this section, we detail three procedures that can be implemented in the programming language `R` (http://www.r-project.org). We decided to present them as three separate procedures as it may be desireable to disentangle them in an application. Note that the first two procedures require data prior to the treatment whereas the third does not. The first procedure `bsc.choice()`, identifies the set of confounders and scale of the outcome variable that minimize the objective function in (9). The second procedure `bsc.test()`, tests if the bias stability condition is violated via (23). The final procedure `npdid.estimation()`, estimates the treatment effect in (14). All `R` code can be requested from the authors upon publication of the article.

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.[43] 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 continous) can vary with each data frame. It is feasible to have some of the confounders in a given set to be 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 where the treatment was administered.[44] This variable 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.[45]

The function consists of several steps. It first determines the type of variable (ordered, factor or continuous) from each 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

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

[44] 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.

[45] 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.

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, (9) is calculated for each set of confounders. The procedure then determines the set which minimize (9).

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

`y`: The outcome variable associated with the smallest value for (9).

`x`: The set of confounders that minimize the objective function.[46]

`bsc.store`: The value produced for each set of confounders of (9).

`min.bsc.store`: The minimum value of produced amongst the set of confounders of (9).

`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 should take the resulting outcome variable and set of confounders and conduct the `bsc.test()` with those variables. We discuss this function in the next subsection.

Description of the function bsc.test()

The main purpose of this function, `bsc.test()`, is to test if there is a violation of the bias stability condition. 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

`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 where the treatment was administered. This variable 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.

[46] 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.

`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, (23) is calculated. A bootstrap[47] is used to them produce 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 (23).

`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.

The figure plots the estimated density of the bootstrapped test statistics[48] 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 $n \times 1$.

[47] 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.

[48] 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")`.

`t`: The time period. This is a discrete variable which must be equal to zero in the period where 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 (34) 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 rate of convergence of the other three groups (treated in period 0, control n period 1 and control in period 0).

The TT is then calculated as in (14). A bootstrap[49] is used to them 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 any of the 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 after publication of the article.

[49]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.