I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15799

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

Daniel J. Henderson
Stefan Sperlich

DECEMBER 2022

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15799

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

Daniel J. Henderson
University of Alabama and IZA

Stefan Sperlich
University of Geneva

DECEMBER 2022

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

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

We propose a complete framework for model-free difference-in-differences analysis with covariates, where model-free means data-driven, in particular nonparametric estimation and testing, variable and scale choice. We start with searching for the preferred data setup by simultaneously choosing confounders and a scale of the outcome variable along identification conditions. The treatment effects themselves are estimated in two steps: first, the heterogeneous effects stratified along the covariates, then the average treatment effect(s) for the population(s) of interest. We provide the asymptotic statistics as well as the finite sample behavior of our methods, and suggest bootstrap procedures to calculate standard errors and p-values of significance tests. The pertinence of our methods is shown with a study of the impact of the Deferred Action for Childhood Arrivals program on human capital responses of non-citizen immigrants. We show that past results underestimated the positive impact on school attendance for individuals aged 14-18, and the positive impact on high school completion. Moreover, we find that the parametric methods fail to identify the negative impact on school attendance of college aged individuals. Practical issues including bandwidth selection, sample weights, and implementation are given in the supplement.

JEL Classification: C14, A2

Keywords: nonparametrics, causal analysis, difference-in-differences estimators, heterogeneous treatment effects

Corresponding author:
Stefan Sperlich
Research Center for Statistics
Geneva School of Economics and Management
Université de Genève
Bd du Pont d'Arve 40
CH-1211 Genève
Switzerland

E-mail: stefan.sperlich@unige.ch

1 | INTRODUCTION

Arguably the most popular estimation technique to study treatment effects in a Rubin-Causal-Model (Holland, 1986) is the so-called difference-in-differences (DiD) approach. This is feasible when via a panel or repeated cross-sections of individuals are observed both before and after an intervention has taken place. Our methods are outlined for repeated cross-sections (cohorts); balanced panels give simplified versions. We call such 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 (treatment group, $D = 1$ and control group, $D = 0$). The primary assumption behind this method for identifying the treatment effect on the treated is that without such intervention, the (conditional) outcome 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. This insinuates that there is only a constant difference between the two groups, disturbed by this treatment.

Often 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, while neither originate from treatment itself. In the former case you can think of an interaction between a (pre-)condition and time, and in the latter of an exogenous change of conditions over time. These fears can be mitigated by proper conditioning, say by including confounders X. While for identification a common trend, conditional or unconditional, is only required for a given period around treatment, it seems reasonable to assume that this should also hold for the period(s) before the intervention. The same could be said about periods after treatment only if the treatment simply shifts the development of Y by a constant.

For the considerations above, we focus on the difference in differences of means, namely

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

where in $E[Y_s|x, d] := E[Y_s|X_s = x, D_s = d]$, we condition the expectation of Y on a set of confounders X and treatment status D in period s. For simplicity we consider $d \in \{0, 1\}$, i.e. treated ($d = 1$) and controls ($d = 0$). When treatment takes place between periods $t-1$ and t, (1) gives the conditional treatment effect on the treated from which we can obtain average effects.[1] This is based on the assumption that (1) had been zero for all x without treatment. Thus, to identify a causal effect, we work with a scale for Y and a set X such that (1) is zero (noting that both choices have consequences for the interpretation). Using this statistic can turn a bane into a boon: while it may be difficult to convince others that this assumption is fulfilled, an appropriate statistic can guide you data-adaptively. Assuming data is available in at least one period prior to treatment (say, $t = -1$), we can check if (1) is zero for a given X prior to treatment (say, the development between $t = -1$ and $t = 0$). We emphasize that while this is not the (non-testable) identification condition needed, it empirically supports its credibility.

Equation (1) is far more useful than being used to estimate an average treatment effect on the treated (TT). We study its estimation, including heterogeneous TT, its sample average (i.e., the TT itself), and the analogue of its squares (i.e., test statistics). In each case, we study the asymptotic and finite sample properties. In practice, it is likely preferable to rely on bootstrap methods than on estimates of complex asymptotics. For the test, a challenge is to find then procedures that generate data under the null hypothesis.

Without confounders, the linear DiD estimator is identical to the nonparametric TT 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

[1] For simplicity, we will consider three time periods $t = -1, 0$ and 1. The treatment will occur between periods 0 and 1.

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. If the unconditional treatment effect is of interest, you need to have more than three continuous variables to be affected asymptotically; else, often higher smoothness conditions are assumed that allow for bias reduction to end up with the optimal rate. Many variables are discrete, and many continuous variables are measured or recorded discretely (e.g., years of education), and a nonparametric approach is reasonable even when all confounders are discretely measured. In our application we show that this holds true even computationally.[2] The case of mainly or even only discrete regressors is surprisingly common. For example, in our data analysis, Kuka et al. (2020) examine human capital responses to the availability of the Deferred Action for Childhood Arrivals (DACA) program. In addition to having all binary right-hand-side variables, their outcome variables are binary. As 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.

Our contribution is the introduction of a complete framework for model-free DiD based causal analysis under the potential presence of confounders. We start by presenting a data-driven procedure to find an appropriate scale of Y with a set of confounders compounded in a vector X that (both together) prove to have some credibility to identify the treatment effects via the 'parallel path'. As this cannot be done for the period of interest itself, we can study the parallel path for previous periods (i.e., not the actual assumption but an indicator for its plausibility). We then estimate the identified effects on the treated. The procedure is concluded by the introduction of nonparametric tests for significant treatment effects. Modified versions of the simultaneous test for significance of conditional effects can be used for testing heterogeneity of effects or the credibility of identification assumptions. The analytical developments (with technical details in the Supplement) are completed by simulations (deferred to the Supplement) which show the usefulness of all methods even for very small samples.

One may ask about post-selection (or pretesting) inference as we propose a procedure that allows you to select between different covariates and scales of Y, or to test for bias stability before treatment started. However, our problem differs from the post-selection inference typically considered (cf. Rolling and Yang (2014) for the treatment effect estimation context). Specifically, Taylor and Tibshirani (2015) describe the standard problem as follows: "Having mined a set of data to find potential associations, how do we properly assess the strength of these associations? The fact that we have cherry-picked, i.e. searched for the strongest associations means that we must set a higher bar for declaring significant the associations that we see." Our criterion is not the covariates contribution to a regression, but the maximization of bias stability, i.e. checking the identifying assumptions necessary for causal conclusions. However, as this is infeasible for the period of interest, it has to be done for a prior period. That is, there is no cherry-picking for significance or finding the strongest treatment effect; we rather do the contrary, maximizing the conditional independence. Doing this for periods prior to the one of interest suggests to apply a strategy equivalent to sample splitting. The existing literature related to our context even advises against conditioning on such pretests (Roth, 2022).

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

[2] It is relatively straightforward to employ parametric or semiparametric versions of our methods. However, these strict parametric assumptions may or may not be validated by domain knowledge like economic theory, but the misspecification of functional forms typically leads to biased and inconsistent estimates.

Section 2 presents the basics of model-free conditional DiD analysis. Section 3 suggests a tool to evaluate the choice of scale and confounders along the 'parallel path' assumption. Section 4 presents the estimator and its asymptotic properties. This is followed by a general format for nonparametric significance tests in Section 5 together with their bootstrap approximation. Section 6 contains our application, and Section 7 concludes. Technical details, simulations, implementation with practical issues like bandwidth choice and sampling weights as well as details on the R procedure code can be found in the Appendix and Supplement.

2 | NONPARAMETRIC DIFFERENCE-IN-DIFFERENCES

Although our main contribution is not a new estimator, but the provision of a complete framework of the causal analysis, we briefly discuss some of the most related literature as long as it comes along with the corresponding asymptotic theory. For a more general discussion, recalling ideas, definitions and assumptions of DiD with confounders we refer to Frölich and Sperlich (2019). For a linear parametric DiD with confounders you can consult Sant'Anna and Zhao (2020) who considered a so-called double robust version, i.e., using propensity score weighting and regression.[3] 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. In nonparametric estimation, both functions are 'correctly specified', and it is not clear if doing both would result in an improvement in practice. However, Kennedy et al. (2017) introduced a special nonparametric doubly robust matching estimator for continuous treatment whose extension to DiD could be interesting. Abadie (2005) and Qin and Zhang (2008) proposed DiD with non- and semiparametric propensity score weighting, respectively. Chan et al. (2016) proposed a more general weighting scheme for matching, but not the entire DiD. As there is no general superiority of propensity score weighting over the matching, we stick to the latter (i.e., a DiD regression approach). The advantage is that then we do not need to jump between nonparametric propensity estimation to nonparametric regression and back. We also avoid numerical problems that occur when dividing by nonparametric estimates of potentially small propensities.

Another reason is that for exploring potential heterogeneity of effects beyond confounders (variables that partly predict both, D and Y), you need to regress on these additional covariates in X anyhow, as (conditional on the confounders) the propensity score does not exhibit any variation on them. As Frölich and Sperlich (2019) discuss, there are further reasons you 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. Their inclusion can improve the statistical analysis by resting noise. Which covariates to choose is seemingly the researcher's choice, but this has implications for both, interpretation and assumptions. As we condition on both, confounders and additional covariates, we will henceforth speak of 'covariates' in general. 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. Before estimating the treatment effect, we should have a closer look to the identification conditions. We need to observe a set of covariates X, and know the scale of Y, such that stochastically speaking, the

[3] This is not to be mixed up with double machine learning or double debiased methods. These are completely different concepts, both designed to tackle problems we don't have.

parallel path and a common support condition holds in the mean. However, as we need to assume the common trend for the period in which the treatment takes place, we have to introduce the notion of 'potential outcomes' for Y, where Y^d represents the response that would be obtained if treatment $D = d$ had taken place. We further need to define the domain $\mathcal{X} \subset supp(X)$ which is implicitly determined by the so-called common support condition (CSC) which says that $\forall x \in \mathcal{X}, \forall (t,d) \in \{(0,0),(1,0),(0,1)\}$

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

where for the sake of notation time T is dealt with like a random variable. Loosely speaking, we assume 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. If necessary, we can redefine the population of interest such that this holds. For identification of the treatment effect, observations before treatment ($t = 0$) are supposed to be free of anticipation effects; else you only measure the treatment minus anticipation. Specifically:

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

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

We are no longer looking for a parallel path of the potential outcomes Y^0, but of $Y^0|x$, an important distinction when switching from unconditional to conditional DiD. Moreover, (2) 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 well-known TT matching estimator. We only assume that this difference is the same before treatment, suggesting that we can use pre-treatment data for bias correction. Therefore, calling Assumption I 'bias stability' is perhaps more appropriate as it does not deceptively insinuate a parallel path of Y^0.

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

Denote the conditional expectations for each year and treatment group by

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

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

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

and consequently also the unconditional TT for any (sub-)population, by integrating out x accordingly. Let n_{dt} denote the number of observations in group d at time t, and suppose that all n_{dt} converge at the same rate to infinity. Further, denote TT_a as the TT that results from integrating TT_x over the distribution of x in the group with $D = T = 1$. 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 will speak of TT when we refer to both. Recall that we do not require a balanced

panel. Consequently, we typically do not observe X for all people at all time points. All our methods and results are applicable to the simpler case of balanced panels. A balanced panel from the onset does not lead to equivalent results for repeated cross-sections, but it simplifies the asymptotics as will be shown further below and in the supplement.

2.2 | Nonparametric conditional expectations

Most empirical papers use linear panel data methods to estimate the TT. While the linear specification without covariates is equivalent to the method derived via conditional expectations, there is no such result here (Meyer, 1995). Even if 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.[4] We are not aware of any practical work having done this; such inclusion 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. Practitioners often ignore the use of these methods and use the curse of dimensionality as their argument against them. Yet, in most common settings, the curse of dimensionality is not an issue as only very few of the covariates are actually continuously measured. This is even more so for the DiD compared to all competitors for nonexperimental data, as the differencing already accounts for many counfounders.

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

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

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 covariateswithout 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 bias equals

$$B_{dt}(x, h, \lambda) \quad = \quad h^2 \left[\nabla^t m_{dt}(x) \nabla f_{dt}(x) / f_{dt}(x) + tr\{\nabla^2 m_{dt}(x)\}\right] \int w(u) u^2 du \tag{7}$$

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

[5] The notation for the bandwidths h and λ are distinct because of the asymptotic properties for continuous vs discrete variables. We do not have a second set of bandwidths (just one per covariate). In practice, one can use a separate bandwidth (i.e., h_l, λ_k) for each covariate. For notational convenience we treat them as equal in our theory ($h_l \; \forall l$, and $\lambda_k = \lambda \; \forall k$). The theoretical extension is straightforward.

$$+\lambda \sum_{\tilde{x}, d_{\tilde{x},x}=1} \left\{ m_{dt}(x^s, \tilde{x}^c) - m_{dt}(x) \right\} f(x^s, \tilde{x}^c) f_{dt}^{-1}(x)$$

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

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 higher smoothness conditions on $m_{dt}(\cdot)$ and the densities of the continuous covariates. In our application, however, all of our covariates are discrete and hence a local-polynomial estimator is not only unnecessary, it is infeasible.

3 | 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 domain knowledge like intuition or economic theory may tell you assuming a common trend is sensible, it does not necessarily tell you the right scale nor the right set of covariates $X^S \subseteq X$ for which it holds. One uses domain knowledge to help specify the causality model, but allow the data to drive the set of confounders, scale of Y and the form of the conditional expectations. For the ease of presentation, we only consider TT_a; modifications for TT_b and TT_x are mostly evident.

The scale matters as it is well known that generally, if the common trend (2) holds for one scale of Y, it can hold for affine, but not for nonlinear transformations due to Jensen's Inequality. The scale of Y is obviously irrelevant for Assumption I if there is no trend or if there is no selection bias (i.e. both sides of (2) are zero); Roth and Sant'Anna (2021) discussed time invariant mixtures of these two cases. For all other situations the scale is important. Unless the researcher has a strong opinion about it, this could be chosen data-adaptively. The covariates are often driven by reasons of total versus partial TT estimation, the reduction of noise, and Assumption I. While the first is fully up to the researcher's interest, the second should be limited to a few cases (due to its implications for interpretation), the third could be done data-adaptively.

Although we propose a feasible, computationally inexpensive procedure, both choice problems are theoretically intertwined. Therefore, the data-adaptive choices should be based on the same objective function and be considered as a simultaneous problem. The objective is to comply with Assumption I, but as all non-treatment outcomes Y^0 are observed only prior to the treatment, we consider periods prior to treatment, e.g.

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

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. One would choose a transformation and covariates that minimize (9). Alternatively, we could likewise integrate (9) over the x_{i1} of the treated in $t = 1$.

As discussed, (9) does not fully correspond to Assumption I, it only gives credibility to it. This is the reason why we speak of evaluation, not of testing. It also has little to do with the typical variable selection problem, especially popular in treatment effect estimation with LASSO. The target in that literature is efficient estimation, while identification is

already taken as granted, and the objective function is a penalized least squares or moment condition for estimation. It has nothing to do with our objective or procedure. Moreover, our above objective function is different from the one we will use below for estimation, as such, popular procedures for debiasing or post-selection inference have no meaning here. Following Kuchibhotla et al. (2022), the only feasible way we see here for addressing the post-selection problem is to perform an analogue to sample-splitting; either to use $t < 0$ in (9), or to split the samples of time point 0 when using $t = 0$ in (9). In the case of facing panel data we still need some orthogonality assumption for the residuals. It is also the potential auto-correlation in the residuals that destroys selective inference in this situation (cf. Roth (2022)). In our conclusions we discuss bootstrap based inference that could account for the variability of our entire procedure. In practice, instead of doing post-section or selective inference, we could do a robustness check by performing estimation and testing not only for the best scale-and-covariates combination found in the prior-to-treatment periods, but also for the second and third best.

3.1 | Data-driven evaluation of potential scales

Finding a strictly monotonous transformation of Y that fulfills (2) corresponds to finding a proper scale. Consequently, it must provide a reasonable interpretation as back-transformation gets affected by Jensen's Inequality. Unless you face one of the discussed situations in which the scale of Y is irrelevant for (2), asymptotically that transformation is unique. For finite samples, however, this does not need to be the case. In practice, this is not an issue as for interpretation reasons we would only compare two to four different scales. You may think of the Box-Cox transformation which depends on a parameter θ giving $Y(\theta)$ but you only consider $\theta \in \{0, 0.5, 1\}$, say from a set Θ. For each set x^S of covariates, there exists a parameter value θ^S_{opt} that optimizes the common trend condition. Clearly, (9) looks at the squared deviations from Assumption I in a prior period, and can thus be understood as a measure of variation. Since variations are scale dependent, we propose to adapt the criterion by accounting for the variance of $Y(\theta)$, and define for any given S and a fixed $t < 1$, the optimal transformation parameter for Y by

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

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

As nonparametric conditional expectation estimators depend on bandwidths, it is worth mentioning that for this step, we do not need optimal bandwidths for each θ. It is sufficient to have a bandwidth for which the selection outcome along the above criterion does not importantly change compared to the outcome based on an optimal bandwidth. This statement can hardly be defined more precisely due to different uncertainties we face, including the variance of various estimators, and the question of how we define 'optimal bandwidth' in our context. In practice, we ask that for the grid of values over which we search for θ, our working bandwidth picks the same θ^S_{opt} (or a very similar one) as the optimal bandwidth would. We suggest using the computationally attractive plug-in bandwidths (see Henderson and Parmeter (2015) and Chu et al. (2015)). For small samples, these tend to slightly oversmooth what would stabilize the numerical performance of the selection procedure. You should not search for the optimal bandwidth using a criterion like (9) or (10) as these criteria are supposed to be based on reasonable estimates of $\widehat{m}(\cdot)$, but not vice-versa.

3.2 | Data-driven evaluation of confounder sets

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

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

defines the optimal set along the analogue to (10), i.e., you jointly calculate the same criterion for all (θ, S) combinations to obtain $(\theta^{S_{opt}}_{opt}, S_{opt})$ which is the most credible regarding the DiD identifiability assumption. In practice these may not be unique for a given data set; then the practitioner may try both but should keep in mind that they may define somewhat different treatment effects.

If initially there are too many sets, one can even perform pre-selection procedures. A simple method is a visual check to see to what extent a covariate could be a confounder. When plotting the distribution of a potential confounder per group and year, these should exhibit different features, either between groups or else between years; otherwise they are not confounders. Certainly, pre-selection could 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, those procedures are more complex than directly applying (11) (Hall et al., 2007). Moreover, they are based on objective functions different from minimizing the deviations in (9). Generally we would advise against mixing different objective functions when the objective is actually the same.

In practice, we suggest using a penalty factor to account for too many covariates. We tried several alternatives, but found that a simple AIC factor worked well in simulations. Considering our criterion in (11), we propose to add

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

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

It is possible to formally conduct a nonparametric significance test to see if Assumption I is rejected for any given pair (θ, S) for the period before treatment (in practice, you would test this at the "optimal set"). This can be done by taking (9) as the test statistic (for $t = 0$), see Section 5 and the Supplement. However, note that you can only test the credibility of Assumption I, not the assumption itself.

4 | TREATMENT EFFECT ESTIMATORS

4.1 | Conditional treatment effect on the treated

To simplify notation, let Y and X now denote the adequately scaled response and the chosen covariates. Define the DiD estimators of conditional TT

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

Recalling Section 2.2, we immediately obtain

Proposition 1 *Under the assumptions (A1) and (A2) of Racine and Li (2004), extended to the four groups, and assuming independence of errors $u_{it} := Y_{it} - m_{dt}(X_{it})$ for all groups, for all x being interior points for each group, $\widehat{TT}_x$ has a smoothing bias which is the difference of differences of the corresponding individual biases given in (7), i.e., $\{B_{11}(x) - B_{01}(x)\} - \{B_{10}(x) - B_{00}(x)\}$. Similarly, its asymptotic variances are the sum of their asymptotic variances, i.e., $\Omega_{11}(x)/(n_{11}h_{11}^p) + \Omega_{01}(x)/(n_{01}h_{01}^p) + \Omega_{10}(x)/(n_{10}h_{10}^p) + \Omega_{00}(x)/(n_{00}h_{00}^p)$. The biases and variances resulting from the smallest n_{dt} will dominate the others. Following (5), $\widehat{TT}_x$ converges at this rate to a normal distribution.*

It is well known that the assumptions could be modified, but for simplicity, we stick with the work of Racine and Li (2004). We allow each bias term to have its own set of bandwidths (h_{dt}, λ_{dt}). As sign and smoothness of the $m_{dt}(\cdot)$ should not change over d and t, equation (7) suggests that the differencing has not only a bias reducing effect regarding identification (i.e., a potential specification bias), but also regarding smoothing.

In the popular setting of balanced panels ($n_1 = n_{11} = n_{10}$, $n_0 = n_{01} = n_{00}$) and conditioning only on covariates from $t = 0$, assuming $u_{i0} \perp u_{i1}$ for all i becomes less credible.[6] However, the asymptotics simplify nonetheless, as now we have for $d = 0, 1$

$$\widehat{m}_{d1}(x) - \widehat{m}_{d0}(x) = \frac{\sum_{D_i=d:i=1}^{n_d} K(X_{i0}, x, h_d, \lambda_d)\,(Y_{i1} - Y_{i0})}{\sum_{D_i=d:i=1}^{n_d} K(X_{i0}, x, h_d, \lambda_d)}. \tag{14}$$

Corollary 2 *For balanced panels with $\tilde{\sigma}_d^2(x) = Var(u_{i1} - u_{i0}|X_{i0} = x, D = d)$, and conditioning only on covariate values observed in $t = 0$, but else the same assumptions as in Proposition 1, the bias expression remains the same, whereas the variance is now $\tilde{\sigma}_1^2(x)/(n_1 h_1^p) \int w^2(v)dv\, f_{10}^{-1}(x) + \tilde{\sigma}_0^2(x)/(n_0 h_0^p) \int w^2(v)dv\, f_{00}^{-1}(x)$.*

Unfortunately, the asymptotics of the unconditional TT are not that straightforward, see below. In practice, no one would try to estimate the bias and variance of $\widehat{TT}_x$, especially not for all potential x. Even the estimation of the variance of $\widehat{TT}_a$ or $\widehat{TT}_b$ can hardly be recommended. Instead, we use a wild bootstrap procedure (see Appendix).

Before we turn to the unconditional treatment effects, it is worth recalling two points. First, looking at conditional treatment effects may be the most insightful way to study (potential) heterogeneity of treatment effects. Therefore we consider the above results not just as an intermediate step for the main result. 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 only 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 a binary variable for sex, you obtain $TT(x_2)$ to study TT for males and females separately.

[6] 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 errors over time are unlikely as we have cohorts.

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 integrating $\widehat{TT}_x$. For the sake of brevity, we consider TT_a, estimated by averaging over the (n_{11}) x_i observed in group $d = 1$ at time period $t = 1$:

$$\widehat{TT}_a = \frac{1}{n_{11}} \sum_{i:\, D_{i1}=1}^{n_{11}} \left\{ \widehat{m}_{11}(x_{i1}) - \widehat{m}_{01}(x_{i1}) - \widehat{m}_{10}(x_{i1}) + \widehat{m}_{00}(x_{i1}) \right\}. \tag{15}$$

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 one only considers 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 CSC is fulfilled, which typically corresponds to trimming at the boundaries. This is convenient for other reasons, like avoiding the necessity of boundary corrections for the estimator $\widehat{m}_{dt}(x)$. To avoid complicating our formulas, we continue with the above notation, assuming that in (15), we only average over interior points.

For the asymptotics, we refer to the fact that in case of independent residuals, statistic (15) can be viewed as an extension of the kernel based matching estimator. It is feasible then to replicate the calculations for nonparametric matching estimators in the existing literature to obtain the bias and variance, and invoke the central limit theorem. The convergence of $\widehat{m}_{dt}(x)$ imply we can choose λ_{dt} and h_{dt} for $dim(X^s) = p \leq 3$ such that $B = o(n_{dt}^{-1/2})$ and $\sqrt{n_{dt} h_{dt}^p} = o(1)$. To achieve this for more than three continuous covariates, we could invoke higher-order kernels or local-polynomial estimators, both based on higher-order smoothness assumptions for $m_{dt}(\cdot)$ and the distributions of X.[7] Asymptotically, for $dim(X^c) = k$, we have no such restriction unless k increases with the sample size.

Proposition 3 *For $p \leq 3$ such that h_{dt}^2 and $n_{dt}^{-2} h_{dt}^{-p}$ are of order $o(n^{-1})$*

$$\sqrt{n_{11}} \left\{ \frac{1}{n_{11}} \sum_{i:\, D_{i1}=1} \widehat{TT}_a(X_{i1}) - TT_a \right\} \to N(0, V_a), \tag{16}$$

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

$$\begin{aligned} V_a \quad = \quad & E\left[\{ m_{11}(X) - m_{10}(X) - m_{01}(X) + m_{00}(X) - TT \}^2 | D = T = 1 \right] \\ & + E\left[\sigma_{11}^2(X) | D = T = 1 \right] + E\left[\frac{\sigma_{10}^2(X) f_{11}^2(X)}{\kappa_{10} f_{10}^2(X)} | D = 1 - T = 1 \right] \\ & + E\left[\frac{\sigma_{01}^2(X) f_{11}^2(X)}{\kappa_{01} f_{01}^2(X)} | D = 1 - T = 0 \right] + E\left[\frac{\sigma_{00}^2(X) f_{11}^2(X)}{\kappa_{00} f_{00}^2(X)} | D = T = 0 \right]. \end{aligned} \tag{17}$$

Uniform rates of convergence could be obtained by following results similar to Racine and Li (2004). In the Appendix we give the influence function (IF) to derive 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

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

estimation for time invariant X, where $D \perp T$, and $D \perp T|X$). From their paper you can also see how our result simplifies for balanced panels.

As the bootstrap inference for these estimators is relatively straight-forward given the existing literature, we deferred it to the Appendix, whereas bootstrap inference for testing is more involved, see below.

5 | TESTING

To complete the cycle of a DiD analysis, we consider some testing problems of interest. We first briefly discuss how to test for significance of an unconditional treatment effect. We then introduce a nonparametric test that can be used to jointly test for the significance of conditional treatment effects as well as for checking if treatment effect heterogeneity is large. It can also be used for supporting Assumption I, recall Section 3.2.

5.1 | Significance of treatment effects

To test for significant treatment effects TT_z of type $z = x,\ a$ or b, we consider hypotheses

$$H_0^z:\ TT_z = 0 \qquad vs. \qquad H_1^z:\ TT_z \neq 0\,. \tag{18}$$

Exploiting (16), we can construct an asymptotic or bootstrap confidence interval (Section 5.4) to see if it includes zero.

5.2 | Composite significance testing in model-free DiD

While Assumption I cannot be directly tested, its credibility can. We can essentially use the same statistic as we did for the selection procedures, namely (9),[8] applied to the pre-treatment period (from $t = -1$ to 0), where by definition, $Y_i = Y_i^0$ for all subjects i.

$$\mathcal{T}_t := \frac{1}{n_{1t}} \sum_{i:D_{1t}=1}^{n_{1t}} \left\{\widehat{m}_{1t}(x_{it}) - \widehat{m}_{0t}(x_{it}) - \widehat{m}_{1(t-1)}(x_{it}) + \widehat{m}_{0(t-1)}(x_{it})\right\}^2, \tag{19}$$

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

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

A bootstrap will require us to resample the data under the null.

Joint significance of heterogeneous 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.

$$\mathcal{T}_1 := \frac{1}{n_{11}} \sum_{i:D_{11}=1}^{n_{11}} \left\{\widehat{m}_{11}(x_{i1}) - \widehat{m}_{01}(x_{i1}) - \widehat{m}_{10}(x_{i1}) + \widehat{m}_{00}(x_{i1})\right\}^2, \tag{20}$$

[8] A rescaling by the response variance estimate is not needed.

is the appropriate test statistic for such null hypothesis (i.e., $M_1(x)$).[9]

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

$$\mathcal{T}_H := \frac{1}{n_{11}} \sum_{i:D_{1t}=1}^{n_{11}} \left\{ \widehat{m}_{11}(x_{i1}) - \widehat{m}_{01}(x_{i1}) - \widehat{m}_{10}(x_{i1}) + \widehat{m}_{00}(x_{i1}) - c \right\}^2 . \tag{21}$$

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.

Bias stability condition To test if the bias stability (parallel path) in the period(s) prior to treatment (e.g., from $t = -1$ to $t = 0$) as an additional check of Assumption I's credibility, we consider the statistic

$$\mathcal{T}_0 := \frac{1}{n_{10}} \sum_{i:D_{10}=1}^{n_{10}} \left\{ \widehat{m}_{10}(x_{i0}) - \widehat{m}_{00}(x_{i0}) - \widehat{m}_{1(-1)}(x_{i0}) + \widehat{m}_{0(-1)}(x_{i0}) \right\}^2 , \tag{22}$$

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

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

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 (21). 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.

Theorem 4 *Define the four one-dimensional densities $f_{dt}(x)$ implicitly by $\int_0^{x_{it}} f_{dt}(x)dx = i/n_{dt}$ for all observed x_{it} with $D_{it} = d$.[10] Assume all $m_{dt}(\cdot)$ and $f_{dt}(\cdot)$ are $r \geq 2$ times continuously differentiable on $[0, 1]$, and kernel $W(X, x, h)$ being of order r. For the optimal testing rate $h = O(n_{11}^{-2/(4r+1)})$ with $n_{11}h^2 \to \infty$, and κ_{dt} as defined after (16), we have*

$$n_{11}\sqrt{h} \left\{ \mathcal{T}_1 - \frac{1}{n_{11}h} \int W^2 \sum_{d,t=0}^{1} \int \frac{\sigma_{dt}^2(x) f_{11}^2(x)}{\kappa_{dt} f_{dt}(x)} dx \right\} \longrightarrow N(0, \mathcal{V}) \quad \text{as all } n_{dt} \to \infty, \tag{23}$$

under H_0, where the variance $\mathcal{V}/(n_{11}^2 h)$ of our statistic $\mathcal{T}_1$ is

$$\frac{2}{n_{11}^2 h} \int (W * W)^2 \left(\sum_{d,t=0}^{1} \int \frac{\sigma_{dt}^4(x) f_{11}^2(x)}{\kappa_{dt}^2 f_{dt}^2(x)} dx + 2 \sum_{mix(dt,ks)} \int \frac{\sigma_{dt}^2(x)\sigma_{ks}^2(x) f_{11}^2(x)}{\kappa_{dt}\kappa_{ks} f_{dt}(x) f_{ks}(x)} dx \right), \tag{24}$$

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

For the case where the statistic $\mathcal{T}_1$ averages over the $n_1 = n_{11} + n_{10}$ treated, replace n_1 for n_{11} and $f_1(\cdot)$ for $f_{11}(\cdot)$

[9] As you may prefer TT_b over TT_a, you can also average in (20) over all treated ($n_{11} + n_{10}$).

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

in (23), (24), and in the definition of κ_{dt}. Its extension to allow for the inclusion of weights and trimming (Supplement).

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 (23) 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 rate. Due to (14), for balanced panels we get

Corollary 5 *Consider a balanced panel taking all covariate values from $t = 0$ with $\tilde{\sigma}_d^2(x) = Var(u_{i1} - u_{i0}|x, D = d)$, and let $f_1(\cdot)$ define the density of X for the treated, $f_0(\cdot)$ for the controls. Then, along with the assumptions from Theorem 4,*

$$n_1\sqrt{h}\left\{\mathcal{T}_1 - \frac{\int W^2}{h}\int \frac{\tilde{\sigma}_1^2(x)f_1(x)}{n_1} + \frac{\tilde{\sigma}_0^2(x)f_1^2(x)}{n_0 f_0(x)}dx\right\} \longrightarrow N(0, \tilde{\mathcal{V}}), \tag{25}$$

under H_0, for $\kappa_d = \lim(n_d/n_1)$, and with

$$\tilde{\mathcal{V}} = 2\int (W * W)^2\left(\sum_{d=0}^{1}\int \frac{\tilde{\sigma}_d^4(x)f_1^2(x)}{\kappa_d^2 f_d^2(x)}dx + 2\int \frac{\tilde{\sigma}_1^2(x)\tilde{\sigma}_0^2(x)f_1(x)}{\kappa_1\kappa_0 f_0(x)}dx\right). \tag{26}$$

5.4 | Feasible bootstrap tests

Arguments in favor of using a bootstrap for testing are as strong as for estimation. We need large samples before the first-order terms fully dominate the second and third-order terms. Even if the samples were large enough to trust the normal approximation, estimation of the first-order terms would still remain a non-trivial problem. 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. Our proposal follows ideas of the related literature, 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. Different scenarios are conceivable to comply with H_0^1. For that reason, we need to take the residuals 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). The steps are:

1. Pool data (over treated and control groups) within each year t, $(t-1)$, and estimate $m_{t=1}(x) := E[Y|t = 1, X = x]$ for all x observed in $t = 1$. Analogously, $m_{t=0}(x) := E[Y|t = 0, X = x]$ for all x observed in $t = 0$.
2. Generate $B \geq 100$ 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 $Y_{it}^{*b} = \widehat{m}_t(X_{it}) + u_{it}^{*b}$, for given d, t, $i = 1, ..., n_{dt}$, where u_{it}^{*b} might be generated by $\widehat{u}_{it}$ times an independent $N(0, 1)$ variable.
3. From these samples, calculate B estimators $\mathcal{T}_1^{*b}$ which are calculated as in (22), 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 has the tendency to suffer from size distortions in the sense of over-rejection. A 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. This was confirmed by many simulations.

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

6 | APPLICATION: HUMAN CAPITAL RESPONSES TO DACA

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

Prior to discussing the application, we mention that we performed intensive simulations beforehand. The performance was checked for all our above presented methods. All details and results are given in the Supplement. We recommend and use practical bandwidth choice procedures, typically plug-in methods, and complete them with implementation instructions and R code. Even for very small data sets, all procedures worked surprisingly accurate and robust in all our simulations. Only for some special situations we found - as expected - problems with the size in more complex testing problems.

6.1 | Data

The data come directly from Kuka et al. (2020) and we only discuss them briefly (the data are freely available on `doi.org/10.1257/pol.20180352`). Kuka et al. (2020) use the Integrated Public Use Microdata Series (IPUMS) American Community Survey (ACS) (Ruggles et al., 2018) over the period 2005–2015. They focus on (a sample of) immigrant youth aged 14 to 22 during the time of the survey such that they arrived on US soil by the age of 10 in 2007. The sample from 14–18 is used to study high school attendance, while the sample from age 19–22 is used to study high school completion (including those who graduated from high school as well as those who earned a passing grade on the General Educational Development test) and post-secondary attendance (three different binary left-hand-side variables). Recall that with a binary outcome, linear DiD estimators do not guarantee the predicted outcome lies between zero and one. Our nonparametric estimator guarantees this support condition.

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

for age of immigration-by-eligibility and age-by-eligibility fixed effects.[11] 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.[12]

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

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

6.2 | Empirical results

The parametric results can be found in Tables 1 and 2. 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 (see the Supplement) and use our bootstrap procedure (with $B = 999$) to calculate our standard errors.

The final three values associated with each sample in Tables 1 and 2 are the sample size, the mean of the outcome variable and the p-value associated with our bias stability test. The latter shows mixed results.[13] In Table 1, we firmly reject the null that the BSC ("parallel path") holds in our sample for 14-18 year olds, but are unable to reject it for each case for 19-22 year olds. Table 2 shows four cases where we fail to reject the null and five cases where we reject the null. As we are simply looking to replicate the results of their paper, we proceed as if we were unable to reject the null hypothesis in each scenario.[14] We therefore should be careful about the interpretation of each treatment effect as identification is in question for several of them. In practice, we would suggest that more potential covariates be tracked down in order to satisfy the identification condition.

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

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

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

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

6.2.1 | School attendance

The results for school attendance are found in Table 1. For individuals aged 14-18, the parametric models show positive and significant estimates for each grouping (all, hispanic and high take-up sample). These results suggest that DACA led to an increase in school attendance of 1.2 percentage points among all immigrants with 2.2 and 2.9 percentage point increases for Hispanic and high take-up sample immigrants.

If we look to the nonparametric results for those aged 14-18, they are larger (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 1 Effect of DACA on school attendance

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

Table 1 also gives the results for 19-22 year olds. While this group was primarily used to examine later schooling outcomes, it is interesting to see these impacts. The parametric model gives positive, but insigificant estimates. The nonparametric model gives negative and significant estimates for each sample. There is substantial evidence in the literature to suggest that the impact of DACA on college age enrollment is in fact negative. Hsin and Ortega (2018) found that DACA increased dropout rates by 7.3% in 2018. Amuedo-Dorantes and Antman (2017) found that DACA reduced the probability of school enrollment of eligible higher-educated individuals as it increased the likelihood of employment of men. The lack of authorization led individuals to enroll in school when working legally was not feasible. While the differences in point estimates with respect to 14-18 year olds is interesting, the ability of our method to identify the negative impact on college-aged individuals shows the downsides of relying on parametric assumptions.

6.2.2 | High school completion and college enrollment

The effects of DACA on high school completion and college enrollment can be found in Table 2. The first three columns represent the effect on high school completion (GED or diploma) for all immigrants, Hispanic immigrants and immigrants from high take-up countries, respectively. These results are broken down by age (19, 19-22 and 23-30). Similarly, the fourth through sixth columns give the impact of DACA on the completion of some college (more than 12 years of education completed) for each of the groups (all, Hispanic, and high take-up) for each age group.

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

TABLE 2 Effect of DACA on high school completion and college enrollment

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

The nonparametric results are equally interesting. Here we find the impact of DACA on high school completion to be larger than that found in Kuka et al. (2020). For 19 year olds, the nonparametric model suggests that the increase was 9.6 percentage points for all immigrants, 12.8 percentage points for Hispanic immigrants and 15.2 percentage

points for immigrants from high take-up countries. That being said, these point estimates are not statistically different from their corresponding parametric counterparts.

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

A similar patter occurs for the impact of DACA on some college. The fourth through sixth columns of Table 2 show higher impacts of DACA in the nonparametric setting (except for the high take-up sample) for 19 and 19-22 year olds and lower impacts of DACA for 23-30 year olds. However, the majority of point estimates here are insignificant. While the nonparametric model removes restrictive assumptions, it is unable to conclude that DACA has a significant impact on college enrollment.

In summary, our 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.[15] 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).

7 | CONCLUSIONS AND DIRECTIONS FOR FURTHER EXTENSIONS

We suggest a complete framework for causal analysis (with covariates) via model-free DiD estimation and testing. We show how to automatically select confounders and the scale of the outcome variable, estimate TTs, choose bandwidths and construct standard errors and confidence intervals. We also present model-free testing for significance and heterogeneity of treatment effects. Importantly, we also provide a bootstrap test for credibility of the identification assumptions. These results can be used in many common situations and result in robust analysis. We provide asymptotic theory for both cohorts and panels, for time-varying and for time constant covariates. The finite sample performance has been verified by simulation studies under rather complex designs.

We apply our techniques to study the impact of DACA on human capital decisions. We compare our results to Kuka et al. (2020). If their models were correctly specified, we would expect that we get similar results. As in their paper, we find a positive (but larger) impact of DACA on high school attendance and high school completion, but we also find that they were unable to identify the negative impact of DACA on school enrollment of college aged individuals. Our findings are closer to what intuition suggests.

We proposed a selection of scale and covariates along (9), (10) and (11) in the spirit of the non-testable identifying Assumption I. If one wants to address the post-selection inference problem, we suggested an equivalent to the sample splitting approach (Kuchibhotla et al., 2022). Alternatively, to account for all variation of the entire statistical analysis, we could apply an outer bootstrap loop that runs over all steps of the analysis until the final estimate. In practice this would be extremely costly and may also give unreasonably large standard errors. In our context (i.e. given the objective of the first steps), it is questionable if the practitioner should be interested in such variance.

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

References

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

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

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

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

Chan, K. C. G., Yam, S. C. P. and Zhang, Z. (2016) Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society, Series B*, **78**, 673–700.

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

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

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

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

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

Holland, P. W. (1986) Statistics and causal inference. *Journal of the American Statistical Association*, **81**, 945–960.

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

Kahn-Lang, A. and Lang, K. (2019) The promise and pitfalls of differences-in-differences: Reflections on 16 and pregnant and other applications. *Journal of Business and Economic Statistics*, **38**, 613–620.

Kennedy, E. H., Ma, Z., McHugh, M. D. and Small, D. S. (2017) Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society, Series B*, **79**, 1229–1245.

Kuchibhotla, A. K., Kolassa, J. E. and Kuffner, T. A. (2022) Post-selection inference. *Annual Review of Statistics and Its Application*, **9**, 505–527.

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

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

Li, Q., Racine, J. and Wooldridge, J. (2009) Efficient estimation of average treatment effects with mixed categorical and continuous data. *Journal of Business and Economic Statistics*, **27**, 206–223.

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

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

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

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

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

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

Qin, J. and Zhang, B. (2008) Empirical-likelihood-based difference-in-differences estimators. *Journal of the Royal Statistical Society, Series B*, **70**, 329–349.

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

Rolling, C. A. and Yang, Y. (2014) Model selection for treatment effects. *Journal of the Royal Statistical Society, Series B*, **76**, 749–769.

Roth, J. (2022) Pretest with caution: Event-study estimates after testing for parallel trends. *AER: Insights*, **4**, 305–322.

Roth, J. and Sant'Anna, P. H. (2021) When is parallel trends sensitive to functional form? *Tech. Rep. 2010.04814*.

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

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

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

Taylor, J. and Tibshirani, R. J. (2015) Statistical learning and selective inference. *PNAS*, **112**, 7629–7634.

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

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

A | APPENDIX

A.1 | Proof for the asymptotics of the test statistics

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

$$\mathcal{T}_1 = \sum_{d,t=0}^{1} \Gamma_{dt} + 2 \sum_{mix(dt,ks)} (-1)^{d+k+t+s} \Gamma_{dt,ks} + o_P\left(\frac{1}{n_{11}\sqrt{h}}\right), \tag{27}$$

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

$$\Gamma_{dt} = \sum_{D_i=d:i=1}^{n_{dt}} \sum_{D_j=d:j=1}^{n_{dt}} \int W_{dt}(x_{it}) W_{dt}(x_{jt}) dF_{11}(x)\ u_{it}u_{jt} \tag{28}$$

$$\Gamma_{dt,ks} = \sum_{D_i=d:i=1}^{n_{dt}} \sum_{D_j=k:j=1}^{n_{ks}} \int W_{dt}(x_{it}) W_{ks}(x_{js}) dF_{11}(x)\ u_{it}u_{js}, \tag{29}$$

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 (27), but suppressing in Γ_{dt} the $\sum_{D_i=d:i=1}^{n_{dt}} \int W_{dt}^2(x_{it}) dF_{11}(x)\ u_{it}^2$. That is, we consider the $\Gamma : dt, ks$ and

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

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

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

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).

A.2 | Influence functions of estimators

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

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

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

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

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

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

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

$$\begin{aligned} V_b &= Var(\widehat{TT}_b) = \frac{1}{n}E\Big[\frac{p(X)}{E^2[D]}\{m_{11}(X)-m_{10}(X)-m_{01}(X)+m_{00}(X)-TT\}^2+\frac{p_{11}(X)}{E^2[DT]}\sigma_{11}^2(X) \\ &\quad +\frac{p_{10}(X)}{E^2[D(1-T)]}\sigma_{10}^2(X)+\frac{p_{01}(X)}{E^2[DT]}\frac{p^2(X)}{\{1-p(X)\}^2}\sigma_{01}^2(X)+\frac{p_{00}(X)}{E^2[D(1-T)]}\frac{p^2(X)}{\{1-p(X)\}^2}\sigma_{00}^2(X)\Big], \end{aligned} \tag{30}$$

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 (30) 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 (30), note that $p_{1t}(x) = p(x)\,Pr(T=t|D=1,x)$ and $p_{0t}(x) = \{1-p(x)\}Pr(T=t|D=0,x)$. Note that the resulting simplified formula of (30) coincides with the efficiency bounds derived in Sant'Anna and Zhao (2020).

A.3 | Bootstrap inference for treatment effect estimator

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

Given consistent nonparametric estimators for the $m_{dt}(x)$, our residuals are given by $\widehat{u}_{it} = Y_{it} - \widehat{m}_{dt}(X_{it})$, $\quad i = 1,...,n_{dt},\ d=0,1,\ t=0,1$.Generate $B \geq 100$ bootstrap samples $\{Y_{it}^{*b}, (D_{it}=d), t, X_{it}\}_{i=1}^{n_{dt}}$, $b=1,...,B$, for all groups

$$Y_{it}^{*b} = \widehat{m}_{dt}(X_{it}) + u_{it}^{*b}\,, d=0,1,\ t=0,1,\ i=1,...,n_{dt}, \tag{31}$$

where u_{it}^{*b} can be generated by $\widehat{u}_{it}$ multiplied by an independent $N(0,1)$ variable (which performed best in our simu-

lations).[16] 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 described in the main document, 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. If you use the local-constant version and face binary responses, as we do in our application, you can generate bootstrap replicates

$$Y_{it}^{*b} := \mathbb{1}\{\widehat{m}_{dt}(X_{it}) > v^b\} \quad , b = 1, ..., B \tag{32}$$

with randomly drawn $v^b \sim U[0,1]$. In our application, we received essentially the same standard errors when applying bootstrap versions of (31) and (32). In more complex cases, a link function is recommended. Then a semiparametric bootstrap can be applied to draw from the conditional distribution defined by this link: Define a distribution with $Y|X = x \sim \mathcal{G}\{\eta(x)\}$; estimate the index function $\eta(x)$ and its conditional expectation by local-likelihood, and draw the bootstrap responses Y_{it}^* from $\mathcal{G}\{\widehat{\eta}(X_{it})\}$.

[16] Other authors favor the Radamacher distribution, though in a quite different context.

SUPPLEMENT FOR ORIGINAL ARTICLE

A complete framework for model-free difference-in-differences estimation

Daniel J. Henderson[1] | **Stefan Sperlich**[2]

[1]Department of Economics, Finance and Legal Studies, University of Alabama. E-mail: djhender@cba.ua.edu

[2]Research Center for Statistics, Geneva School of Economics and Management, University of Geneva. E-mail: stefan.sperlich@unige.ch

Correspondence
Stefan Sperlich, Research Center for Statistics, Geneva School of Economics and Management, Université de Genève. Bd du Pont d'Arve 40, CH-1211 Genève. Switzerland.
Email: stefan.sperlich@unige.ch

Funding information
Swiss National Science Foundation for the project 200021-192345.

This document contains supplementary material to the main article. It provides the outcomes of intensive simulation studies for all steps of our causal data analysis. It further gives details on the implementation of all steps, including bandwidth selection, sampling weights inclusion, algorithms with a discussion of parametric and semiparametric versions. And it finally includes a description of functions for our R code.

1 | BOOTSTRAP INFERENCE FOR TREATMENT EFFECT ESTIMATOR

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

In cases such as ours, bootstrap is a widely accepted remedy. It is well known (Mammen, 1992), for nonparametric methods, that the naive bootstrap is insufficient (yields inconsistent estimators for most situations), while the wild bootstrap works. Abadie and Imbens (2008) confirmed the failure of naive bootstrap for kNN matching. Politis (2013) emphasized the superiority of nonparametric (which can be seen as a particular version of the wild) bootstrap for model-free prediction. Bodory et al. (2020) studied explicitly the consistency of the wild bootstrap for nonparametric matching estimators.

The distinction between wild and nonparametric bootstrap often reduces to the question of how many moments are asymptotically matched. While asymptotic theory tells us that, the higher the bootstrap residuals match the moments of the original residuals, the more efficient is the procedure, Davidson and Flachaire (2008) argue that you

need quite large samples before this finding becomes effective. Following their recommendations, we propose a simple version (modifications towards higher-moment matching bootstraps are straightforward), first for continuous responses, then for discrete ones.

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

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

Generate $B \geq 100$ bootstrap samples $\{Y_{it}^{*b}, (D_{it} = d), t, X_{it}\}_{i=1}^{n_{dt}}$, $b = 1, ..., B$, for all groups

$$Y_{it}^{*b} = \widehat{m}_{dt}(X_{it}) + u_{it}^{*b} \,, d = 0, 1, \; t = 0, 1, \; i = 1, ..., n_{dt}, \tag{2}$$

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).[1] 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 described in the main document, 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. If you use the local-constant version and face binary responses, as we do in our application, you can generate bootstrap replicates

$$Y_{it}^{*b} := \mathbb{1}\{\widehat{m}_{dt}(X_{it}) > v^b\} \quad , \; b = 1, ..., B \tag{3}$$

with randomly drawn $v^b \sim U[0, 1]$. In our application, we received essentially the same standard errors when applying bootstrap versions of (2) and (3).

In more complex cases, a link function is recommended. Then a semiparametric bootstrap can be applied to draw from the conditional distribution defined by this link: Define a distribution with $Y|X = x \sim \mathcal{G}\{\eta(x)\}$; estimate the index function $\eta(x)$ and its conditional expectation by local-likelihood, and draw the bootstrap responses Y_{it}^* from $\mathcal{G}\{\widehat{\eta}(X_{it})\}$.

2 | INFLUENCE FUNCTIONS

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

$$\begin{aligned} \varphi_a(X) = \tfrac{DT}{E[DT]} \left[m_{11}(X) - m_{10}(X) - \{m_{01}(X) - m_{00}(X)\} - TT_a \right] \\ + \tfrac{DT}{E[DT]} \{Y - m_{11}(X)\} - \tfrac{D(1-T)}{E[DT]} \tfrac{p_{11}(X)}{p_{10}(X)} \{Y - m_{10}(X)\} \\ - \tfrac{(1-D)T}{E[DT]} \tfrac{p_{11}(X)}{p_{01}(X)} \{Y - m_{01}(X)\} + \tfrac{(1-D)(1-T)}{E[DT]} \tfrac{p_{11}(X)}{p_{00}(X)} \{Y - m_{00}(X)\} + R_{h,n_{11}}(X), \end{aligned} \tag{4}$$

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

[1] Other authors favor the Radamacher distribution, though in a quite different context.

$n_{11} = n\,E[DT]$, we immediately get the seemingly simpler variance representation (cf., Proposition 2)

$$\begin{aligned} V_a &= \frac{1}{E[DT]} E\Big[p_{11}(X)\{m_{11}(X) - m_{10}(X) - m_{01}(X) + m_{00}(X) - TT\}^2 \\ &\quad + p_{11}(X)\sigma^2_{11}(X) + \frac{p^2_{11}(X)}{p_{10}(X)}\sigma^2_{10}(X) + \frac{p^2_{11}(X)}{p_{01}(X)}\sigma^2_{01}(X) + \frac{p^2_{11}(X)}{p_{00}(X)}\sigma^2_{00}(X)\Big]. \end{aligned} \tag{5}$$

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

$$\begin{aligned} \varphi_b(X) = \tfrac{D}{E[D]}\,[m_{11}(X) - m_{10}(X) - \{m_{01}(X) - m_{00}(X)\} - TT] \\ + \tfrac{DT}{E[DT]}\{Y - m_{11}(X)\} - \tfrac{D(1-T)}{E[D(1-T)]}\{Y - m_{10}(X)\} \\ - \tfrac{(1-D)T}{E[DT]}\,\tfrac{p(X)}{1-p(X)}\{Y - m_{01}(X)\} + \tfrac{(1-D)(1-T)}{E[D(1-T)]}\,\tfrac{p(X)}{1-p(X)}\{Y - m_{00}(X)\} + R_{h,n_1}(X) \quad . \end{aligned} \tag{6}$$

Consequently, in Proposition 2, $n_1 = n_{11} + n_{10}$ replaces n_{11} and the variance becomes

$$\begin{aligned} V_b &= Var(\widehat{TT}_b) = \frac{1}{n} E\Big[\frac{p(X)}{E^2[D]}\{m_{11}(X) - m_{10}(X) - m_{01}(X) + m_{00}(X) - TT\}^2 \\ &\quad + \frac{p_{11}(X)}{E^2[DT]}\sigma^2_{11}(X) + \frac{p_{10}(X)}{E^2[D(1-T)]}\sigma^2_{10}(X) \\ &\quad + \frac{p_{01}(X)}{E^2[DT]}\,\frac{p^2(X)}{\{1-p(X)\}^2}\sigma^2_{01}(X) + \frac{p_{00}(X)}{E^2[D(1-T)]}\,\frac{p^2(X)}{\{1-p(X)\}^2}\sigma^2_{00}(X)\Big] \end{aligned} \tag{7}$$

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 (7) 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 (7), 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 (7) coincides with the efficiency bounds derived in Sant'Anna and Zhao (2020).

3 | PROOF FOR THE ASYMPTOTICS OF THE TEST STATISTICS

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

$$\mathcal{T}_1 = \sum_{d,t=0}^{1} \Gamma_{dt} + 2 \sum_{mix(dt,ks)} (-1)^{d+k+t+s}\Gamma_{dt,ks} + o_P\left(\frac{1}{n_{11}\sqrt{h}}\right), \tag{8}$$

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

$$\Gamma_{dt} = \sum_{D_i=d:i=1}^{n_{dt}} \sum_{D_j=d:j=1}^{n_{dt}} \int W_{dt}(x_{it}) W_{dt}(x_{jt}) dF_{11}(x)\; u_{it}u_{jt} \tag{9}$$

$$\Gamma_{dt,ks} = \sum_{D_i=d:i=1}^{n_{dt}} \sum_{D_j=k:j=1}^{n_{ks}} \int W_{dt}(x_{it}) W_{ks}(x_{js}) dF_{11}(x)\ u_{it}u_{js}, \tag{10}$$

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 (8), 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 $\mathcal{V}/(n_{11}^2 h)$ with the sum over the four groups. The errors u_{it} belonging to group (dt) are independent not only within this group, but also from those of any other group (ks); all additive terms in $\Gamma_{dt,ks}$ are independent from each other. Taking expectation, the second part of $\mathcal{V}/(n_{11}^2 h)$ containing all mixtures $mix(dt, ks)$ is

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

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).

4 | 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:

$$D_{it} = 1\{0.75X_{it,1} - 0.5X_{it,2}^2 > \epsilon_{it}\} \tag{11}$$

$$Y_{it} \quad = \quad 1 + t(2 + X_{it,1} + X_{it,2}^2) + D_{it} + D_{it} 1\{t \geq 1\} + u_{it} \tag{12}$$

where the treatment effect on the treated is the coefficient on the interaction term (i.e., $TT = 1.0$) in (12).[2] In (12) 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 our criterion function), and the smallest sample size over all n_{dt} ($d \in \{0, 1\}, t \in \{-1, 0, 1\}$).[3] 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.

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.[4]

4.1 | Choice of the confounder set

To see if our method appropriately picks the correct set of covariates, we generate our data as in (12). 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 $\left(2(k+p)^2 + 2(k+p)\right)/(n_{1\bullet} - (k+p))$ does its job.

If we choose at random, then the fraction correctly specified should be approximately 1/3 and if we choose

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

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

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

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.[5] 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.

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

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

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

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

4.2 | 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 1). We conduct our simulations along the problem of studying

[5]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 $\sigma_u^2 = 2.0$.

the bias stability ('parallel path') condition.[6] We generate our data as in (12) to determine the size of the test. To determine the power, we change the indicator function to $1\,(t \geq 0)$ in (12) as this will generate a situation in which the bias stability condition is violated. We again use $n = \Sigma_{t=-1}^{1} n_t = \Sigma_{d=0}^{1}\Sigma_{t=-1}^{1} n_{dt} = 100, 200, 400$, and 800 total observations and estimate the size (and power) of the test at each of the common (arbitrary) values (1, 5, and 10%).

Inference with nonparametric estimation methods can be notoriously difficult. Using the asymptotic variances of tests are often useless and bootstrap procedures can bring large improvements. That being said, it is common to oversmooth with such tests when using the bootstrap. As we mentioned in the main document, we recommend a common approach of oversmoothing when calculating the residuals which are used in the bootstrap procedure (Vilar and Vilar, 2012). We calculate the test statistic ($\mathcal{T}_0$) as outlined in the main document, but calculate the residuals using the bandwidth procedure of Vilar and Vilar (2012).[7] 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).

Remark: 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.

The results for both the size and power of our test ($\mathcal{T}_0$) 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.

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.

4.3 | Performance of 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).[8] 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

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

[7] 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.

[8] 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.

TABLE 2 Size and power of our bias stability ('parallel path') condition test ($\mathcal{T}_0$): The probability of rejection at each significance level (1, 5 and 10%) using $B = 999$ bootstrap replications in each of our 999 simulations, average sample size (to the nearest integer) for n_{10} and $\min(n_{dt})$ given for each overall sample size ($n = \Sigma_{t=-1}^{1} n_t = \Sigma_{d=0}^{1}\Sigma_{t=-1}^{1} n_{dt}$)

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

Section 1. We do not require a bandwidth g and use h for both estimation and in our bootstrap.[9]

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

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

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).[10] The average mean square error (AMSE) also tends towards zero (evidence that our estimator is

[9] 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.

[10] 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

consistent). As expected, the moderate signal-to-noise ratio results in smaller AMSE values for any given sample size (it does not significantly impact the bias). The third column of numbers 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.

5 | 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.

5.1 | 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.[11] 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.,

$$LSCV(h,\lambda) = \sum_{i:D_{i1}=1}^{n_{11}} \left(Y_i - \widehat{E}_{-i}\left[Y_i|X = x_i\right]\right)^2, \tag{13}$$

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).

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 (13) once the selection

deduced from the average over the linear combinations of individual biases $B_{dt}(x,\lambda,h)$, which in turn depends on the particular bandwidth choices, true densities and functions. Importantly, it is minor in size and rapidly converges to zero.

[11] We tried this in our application and found similar results.

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 5.3, 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 discussed in the main document, 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 in the main document, we do not face the choice of h versus g in our test statistics.

5.2 | Sampling weights

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

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

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

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

which, unfortunately, is not common in canned statistical packages. One way to implement this is via the `npksum` tool in the `np package` in R (Hayfield and Racine, 2008). This allows us to calculate $\sum_{i=1}^{n} Y_i w_i K(X_i, x, h, \lambda)$ 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).

5.3 | 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 Assumption I of the main document. 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.[12]

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 $d = 0, t = -1$).[13]

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

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

6 For the combination of (transformation of) Y and (set of covariates) X^S that minimizes the bias stability condition, use a CV routine to best estimate the conditional expectation $m_{dt}(\cdot)$ for the setting $d = 1$ and $t = 1$. Use the scale factors from this setting to select the bandwidths for the conditional expectations for the other three cases ($d = 1, t = 0$, $d = 0, t = 1$ and $d = 0, t = 0$).[14]

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 1 and perform the tests of interest.

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 Section 3 of the main document to select or rank the scales (or transformations) of Y and the sets of covariates X. 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 6.

5.4 | 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 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.[15]

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

[13] 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 $t = -1$.

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

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

6 | 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(s) in Section 3 of the main document. The second procedure `bsc.test()`, tests if the bias stability condition is violated via test statistic $\mathcal{T}_t$ of the main document (Section 5). The final procedure `npdid.estimation()`, estimates the treatment effect $\widehat{TT}_a$. 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.[16] 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.[17] 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.[18]

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 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

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

[17] 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.

[18] 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.

period 0 and control before period 0). Once these are obtained for each set of confounders,

$$\frac{1}{n_{1\bullet}} \sum_{i:\, D_{i\bullet}=1} \left\{ m_{1t}(x_{i\bullet}) - m_{0t}(x_{i\bullet}) - m_{1(t-1)}(x_{i\bullet}) + m_{0(t-1)}(x_{i\bullet}) \right\}^2, \tag{14}$$

is calculated for each set of confounders. The procedure then determines the set which minimize (14).

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

`y`: The outcome variable associated with the smallest value for (14).
`x`: The set of confounders that minimize the objective function.[19]
`bsc.store`: The value produced for each set of confounders of (14).
`min.bsc.store`: The minimum value of produced amongst the set of confounders of (14).
`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.
`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

[19] 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.

period 0 and then adjusts for the rate of convergence of the other three groups (treated before period 0, control in period 0 and control before period 0). Once this is obtained for the set of confounders, $\mathcal{T}_t$ is calculated. A bootstrap[20] 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 $\mathcal{T}_t$.
`sd.bsc`: The standard deviation (standard error of the test statistic) of the bootstrapped estimates of the test statistic.
`p.value`: The p-value associated with the test statistic. This is calculated as the percentage of bootstrapped test statistics which are larger than the original test statistic.

The figure plots the estimated density of the bootstrapped test statistics[21] along with the value of the test statistic itself as a vertical line. If the vertical line does not appear present in the figure, it is likely far to the right which would suggest rejecting the null hypothesis (i.e., a p-value near zero).

| Description of the function npdid.estimation()

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

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

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

`y`: The outcome variable, which is a $n \times 1$ matrix. It contains the outcome variable for each individual in each time period. It may be discrete or continuous.
`x`: The set of confounders, which is a data frame. This is a $n \times q$ matrix where q refers to the total number of confounders.
`d`: The treatment status. This is a binary variable of dimension $n \times 1$.
`t`: The time period. This is a discrete variable which must be equal to zero in the period 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

[20] 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.

[21] 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")`.

convergence of the other three groups (treated before period 0, control in period 0 and control before period 0). The LSCV procedure defined in (13) 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 estimated as outlined in the main document ($\widehat{TT}_a$). A bootstrap[22] 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.

References

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

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

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

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

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

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

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

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

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

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

[22] 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.

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

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

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

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

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

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

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

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

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

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

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