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

Other titles in Foundations and Trends® in Econometrics

Factor Extraction in Dynamic Factor Models: Kalman Filter Versus Principal Components
Esther Ruiz and Pilar Poncela
ISBN:978-1-63828-096-5

Performance Analysis: Economic Foundations and Trends
Valentin Zelenyuk
ISBN: 978-1-68083-866-4

Experimetrics: A Survey
Peter G. Moffatt
ISBN: 978-1-68083-792-6

Climate Econometrics: An Overview
Jennifer L. Castle and David F. Hendry
ISBN: 978-1-68083-708-7

Foundations of Stated Preference Elicitation: Consumer Behavior and Choice-based Conjoint Analysis
Moshe Ben-Akiva, Daniel McFadden and Kenneth Train
ISBN: 978-1-68083-526-7

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

Daniel J. Henderson
University of Alabama
daniel.henderson@ua.edu

Stefan Sperlich
University of Geneva
stefan.sperlich@unige.ch

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Econometrics

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

D.J. Henderson and S. Sperlich. *A Complete Framework for Model-Free Difference-in-Differences Estimation*. Foundations and Trends® in Econometrics, vol. 12, no. 3, pp. 232–323, 2023.

ISBN: 978-1-63828-277-8

Foundations and Trends® in Econometrics
Volume 12, Issue 3, 2023

Editorial Board

Editorial Scope

Topics

Foundations and Trends® in Econometrics publishes survey and tutorial articles in the following topics:

- Econometric Models
- Simultaneous Equation Models
- Estimation Frameworks
- Biased Estimation
- Computational Problems
- Microeconometrics
- Treatment Modeling
- Discrete Choice Modeling
- Models for Count Data
- Duration Models
- Limited Dependent Variables
- Panel Data
- Time Series Analysis
- Latent Variable Models
- Qualitative Response Models
- Hypothesis Testing
- Econometric Theory
- Financial Econometrics
- Measurement Error in Survey Data
- Productivity Measurement and Analysis
- Semiparametric and Nonparametric Estimation
- Bootstrap Methods
- Nonstationary Time Series
- Robust Estimation

Information for Librarians

Foundations and Trends® in Econometrics, 2023, Volume 12, 4 issues. ISSN paper version 1551-3076. ISSN online version 1551-3084. Also available as a combined paper and online subscription.

Contents

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

Daniel J. Henderson[1] and Stefan Sperlich[2]

[1] *Department of Economics, Finance and Legal Studies, University of Alabama, USA; daniel.henderson@ua.edu*
[2] *Geneva School of Economics and Management, University of Geneva, Switzerland; stefan.sperlich@unige.ch*

ABSTRACT

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

Daniel J. Henderson and Stefan Sperlich (2023), "A Complete Framework for Model-Free Difference-in-Differences Estimation", Foundations and Trends® in Econometrics: Vol. 12, No. 3, pp 232–323. DOI: 10.1561/0800000046.

1

Introduction

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

In our opinion, the most intensive and extensive discussion on this topic was provided by Lechner (2011). He showed that the basic concept for identifying the causal effect via DiD applies to more complex situations than previously considered. In this monograph we limit our considerations to the case of a single treatment and two groups (treat-

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

ment group, $D = 1$ and control group, $D = 0$); extensions as discussed by him work in principle the same way.

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

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

or kurtosis, we would either directly compare the entire distributions of both groups before and after the treatment or estimate these higher moments. As for the latter, the general procedure and ideas remain basically the same; we concentrate here on the estimation of the average treatment effect on the treated.

1.1 Central Equation

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

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

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

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

For simplicity, we will mostly assume we have data on three time periods $t = -1, 0$ and 1 and consider the case where the treatment occurs between periods 0 and 1. Given that we have data in an additional period to treatment ($t = -1$), we can check if (1.1) is zero for a given X prior to treatment (the development between $t = -1$ and $t = 0$).

We emphasize that while this is not the (non-testable) identification condition needed, it empirically supports its credibility.

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

1.2 What Does it Mean to be Model-Free?

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

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

statistical model for estimation and testing. When the latter is done nonparametrically, we speak of nonparametric identification of causality since it does not depend on the parametric specification.

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

In practice, many variables can be discrete, and many continuous variables are measured or recorded discretely (e.g., years of education). For this reason it is often argued in the applied economics literature that parametric methods would be sufficient almost always, as we could construct a saturated model. We will later discuss why this is rarely the case (Appendix B.1). Therefore we argue that if most applications contain a continuous covariate or discrete ones with many values, nonparametric methods are the better option for causal analysis in most cases. In fact,

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

nonparametric regression is at least as reasonable as a parametric one even when all confounders are discretely measured. Moreover, in our application we show that this also holds true computationally. We should mention here that it is relatively straightforward to employ parametric or semiparametric versions of our methods if desired. However, those strict parametric assumptions may or may not be justified by prior knowledge like economic theory, and a misspecification of functional forms easily leads to biased and inconsistent estimates.

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

1.3 Structure of the Monograph

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

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

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

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

school attendance of college aged individuals (19-22). With respect to enrolling in college, we can confirm that these effects are insignificant.

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

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

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

Acknowledgements

The authors wish to thank Brant Callaway, Traviss Cassidy, Bill Greene, Sukjin Han, Guido Imbens, Andrew Johnston, James MacKinnon, Chris Parmeter, Jeff Racine, Pedro Sant'Anna, Gonzalo Vazquez-Bare, Le Wang and Jun Zhao for useful comments and suggestions. The authors would also like to thank participants of the following (1) seminars: McMaster University, University Carlos III of Madrid, University of Alabama, University of California, Riverside, and University of Passau and (2) conferences: North American Summer Meetings of the Econometric Society (Montreal), Asian Meeting of the Econometric Society (Miri), International Association of Applied Econometrics Conference (Rotterdam), International Conference on Computational and Financial Econometrics (London), Meeting of the Economics of Education Association (Porto), New York Camp Econometrics (Lake Placid), and West Indies Economic Conference (Mona). All `R` code is available from the authors' upon request and can be used with the understanding that those who use the code will cite this work in any publication which uses the code.

Appendices

A

Proofs

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

A.1 Asymptotics of the Test Statistics

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

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

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

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

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

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

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

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

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

$$\frac{f_{11}(x_{it})f_{11}(x_{js})u_{it}^2u_{js}^2}{f_{dt}^2(x_{it})f_{ks}^2(x_{js})}\Bigg]$$

$$= \frac{1}{n_{dt}n_{ks}h^2}E\Bigg[\left(W*W\left(\frac{x_{it}-x_{js}}{h}\right)\right)^2 \times \frac{f_{11}(x_{it})f_{11}(x_{js})\sigma_{dt}^2(x_{it})\sigma_{ks}^2(x_{js})}{f_{dt}^2(x_{it})f_{ks}^2(x_{js})}\Bigg],$$

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

A.2 Influence Functions

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

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

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

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

Noting that $n_{11} = n\ E[DT]$, we immediately get the seemingly simpler (compared to the one given in Proposition 2.2) variance representation

$$Var(\widehat{TT}_a) = E\Bigg[\Big\{\{m_{11}(X) - m_{10}(X) - m_{01}(X) + m_{00}(X) - TT_a\}^2$$

$$+\sigma_{11}^2(X) + \tfrac{p_{11}(X)}{p_{10}(X)}\sigma_{10}^2(X) + \tfrac{p_{11}(X)}{p_{01}(X)}\sigma_{01}^2(X) + \tfrac{p_{11}(X)}{p_{00}(X)}\sigma_{00}^2(X)\Big\}\tfrac{p_{11}(X)}{E[DT]}\Bigg]\tfrac{1}{n_{11}}.$$

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

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

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

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

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

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

Let us consider the special case of the simplified variance for balanced panels with all covariate values fixed to the observations in $t = 0$, cf. Corollary 2.2 also for notation. It is not hard to see that it can be written along the above expressions as

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

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

B

Additional Discussion and Final Thoughts

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

B.1 Nonparametric versus Parametric Saturated Models

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

re-definition is needed. We would then call any model saturated if it reproduces the same fit as the overparametrized one.

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

Some people switch to what is sometimes also called a 'reasonably' saturated model, which is even less clearly defined. In practice, its choice is either subjective or random; in either case it risks approximation bias which to some extent corresponds to the smoothing bias in nonpara-

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

metrics. The advantage in nonparametrics is threefold then: (1) this choice corresponds to the bandwidth choice and can easily be done in a data-driven way, (2) we understand the risk and know the smoothing bias so that we can deal with it, and (3) computationally it is essentially always feasible as we do not need to split the categorical variables into dummies.

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

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

Regarding the convergence rate: unless your parametric model is correctly specified, a 'reasonably' saturated model requires you to increase, for increasing sample size, the cells generated by a continuous covariate (or by a discrete covariate with infinite support). The optimal

[2]Remember that your estimate is just a realized random variable; unbiasedness only says that the average of a many those estimates converges to the true value, but the mean square error approach aims on minimizing the distance of your estimate to the true value in probability. Moreover, in the parametric world, 'unbiasedness' only means to have such convergence towards the projection of the real world on your model which can be biased or even meaningless; we know nothing about the distance to the 'truth'.

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

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

B.2 Concluding Remarks

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

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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