UNIVERSITÉ DE FRIBOURG
UNIVERSITÄT FREIBURG

WORKING PAPERS SES

An introduction to flexible methods for policy evaluation

Martin Huber

N. 504
VIII.2019

Faculté des sciences économiques et sociales
Wirtschafts- und sozialwissenschaftliche Fakultät

An introduction to flexible methods for policy evaluation

Martin Huber

University of Fribourg, Dept. of Economics

Abstract: This chapter covers different approaches to policy evaluation for assessing the causal effect of a treatment or intervention on an outcome of interest. As an introduction to causal inference, the discussion starts with the experimental evaluation of a randomized treatment. It then reviews evaluation methods based on selection on observables (assuming a quasi-random treatment given observed covariates), instrumental variables (inducing a quasi-random shift in the treatment), difference-in-differences and changes-in-changes (exploiting changes in outcomes over time), as well as regression discontinuities and kinks (using changes in the treatment assignment at some threshold of a running variable). The chapter discusses methods particularly suited for data with many observations for a flexible (i.e. semi- or nonparametric) modeling of treatment effects, and/or many (i.e. high dimensional) observed covariates by applying machine learning to select and control for covariates in a data-driven way. This is not only useful for tackling confounding by controlling for instance for factors jointly affecting the treatment and the outcome, but also for learning effect heterogeneities across subgroups defined upon observable covariates and optimally targeting those groups for which the treatment is most effective.

Keywords: Policy evaluation, treatment effects, machine learning, experiment, selection on observables, instrument, difference-in-differences, changes-in-changes, regression discontinuity design, regression kink design.

JEL classification: C21, C26, C29.

The author is grateful to Colin Cameron, Selina Gangl, Michael Knaus, Henrika Langen, and Michael Lechner for their valuable comments. Address for correspondence: Martin Huber, University of Fribourg, Bd. de Pérolles 90, 1700 Fribourg, Switzerland; martin.huber@unifr.ch.

1 Introduction

The last decades have witnessed important advancements in policy evaluation methods for assessing the causal effect of a treatment on an outcome of interest, which are particularly relevant in the context of data with many observations and/or observed covariates. Such advancements include the development or refinement of quasi-experimental evaluation techniques, estimators for flexible (i.e. semi- or nonparametric) treatment effect models, and machine learning algorithms for a data-driven control for covariates in order to tackle confounding, learn effect heterogeneities across subgroups and target groups for which the treatment is most effective. Policy evaluation methods aim at assessing causal effects despite the problem that for any subject in the data, outcomes cannot be observed at the same time in the presence and absence of the treatment. As an illustration of this fundamental problem for causality, consider the treatment effect of a job application training for jobseekers on employment. Identifying this effect on the individual level requires comparing the employment state for a specific subject at a particular point in time with and without training participation. However, at a specific point in time, an individual can be observed to have either participated or not participated in the training, but not both. Therefore, treatment effects remain unidentified on the individual level without strong assumptions.

Formally, denote by D a binary treatment, such that $D = 1$ if for instance someone participates in a training and $D = 0$ otherwise. Furthermore, denote by Y the observed outcome, e.g. employment. Following Rubin (1974), let $Y(1)$ and $Y(0)$ denote the potential outcomes a subject would realize if D was set to 1 and 0, respectively, e.g. the potential employment state with and without training. It is assumed throughout that $Y(1)$ and $Y(0)$ only depend on the subject's own treatment and not on the treatment values of other subjects, which is known at the 'Stable Unit Treatment Value Assumption', see Rubin (1990). Observed employment Y corresponds to either $Y(1)$ if the individual receives the training ($D = 1$) or to $Y(0)$ otherwise. The fact that not both

potential outcomes are observed at the same time is formally expressed in the following equation:

$$Y = Y(1) \cdot D + Y(0) \cdot (1 - D). \tag{1}$$

It is easy to see that (1) is equivalent to $Y = Y(0) + D \cdot [Y(1) - Y(0)]$, where the observed outcome is the sum of the potential outcome without intervention and D times $Y(1) - Y(0)$, i.e. the causal effect of D on Y. As either $Y(1)$ or $Y(0)$ is unknown depending on the value of D, the treatment effect can in general not be identified for any subject.

Under specific assumptions, however, aggregate treatment effects are identified based on groups of individuals receiving and not receiving the treatment. Two parameters that have received substantial attention are the average treatment effect (ATE, denoted by Δ) in the population, e.g. among all jobseekers, and the treatment effect on the treated population (ATET, denoted by $\Delta_{D=1}$), e.g. among training participants:

$$\Delta = E[Y(1) - Y(0)], \quad \Delta_{D=1} = E[Y(1) - Y(0)|D = 1]. \tag{2}$$

One assumption yielding identification is statistical independence of treatment assignment and potential outcomes. Formally,

$$\{Y(1), Y(0)\} \perp D, \tag{3}$$

where '$\perp$' denotes statistical independence. (3) implies that there exist no variables jointly affecting the treatment and the potential outcomes. It is satisfied by design in experiments where the treatment is randomized, i.e. not a function of any observed or unobserved characteristics like education, gender, or income. The ATE is then identified by the mean difference in observed outcomes across treated and nontreated groups. This follows from the fact that by (1), $E[Y|D = 1] = E[Y(1)|D = 1]$ and $E[Y|D = 0] = E[Y(0)|D = 0]$, while it follows from (3) that $E[Y(1)|D = 1] = E[Y(1)]$ and $E[Y(0)|D = 0] = E[Y(0)]$. As the average outcomes among treated and

nontreated are representative for the respective mean potential outcomes under treatment and nontreatment in the population, $E[Y|D=1] - E[Y|D=0] = \Delta$.

When the treatment is not randomized, however, a mean comparison of treated and nontreated outcomes is generally biased due to selective treatment take-up, implying that subjects in the treated and nontreated groups differ in characteristics that also affect the outcome. Jobseekers attending a job application training could, for instance, on average have a different level of labor market experience or education than those not participating. Differences in the observed outcomes of treated and nontreated subjects therefore not exclusively reflect the treatment effect, but also the effects of such characteristics, which are thus confounders of the treatment-outcome relation. Formally, the selection biases for the ATE and ATET are given by

$$\begin{aligned} E[Y|D=1] - E[Y|D=0] - \Delta &= E[Y|D=1] - E[Y(1)] + E[Y(0)] - E[Y|D=0], \quad (4) \\ E[Y|D=1] - E[Y|D=0] - \Delta_{D=1} &= E[Y(0)|D=1] - E[Y|D=0]. \end{aligned}$$

Different strategies have been developed for avoiding or tackling selection into treatment in order to identify causal effects. This chapter reviews the most prominent approaches, focusing on methods for flexible model selection and estimation particularly appropriate in big data contexts. Section 2 covers methods relying on selection-on-observables assumptions, implying that observed preselected covariates are sufficient to control for characteristics jointly affecting the treatment and the potential outcomes. Section 3 discusses practical issues to be verified in the data when invoking the selection-on-observables assumption, e.g. the similarity of treated and nontreaded subjects used for estimation in terms of observed characteristics, as well as extensions e.g. to multivalued treatments and different treatment parameters. Section 4 covers causal machine learning, where observed covariates are not preselected, but it is assumed that important confounders can be controlled for in a data-driven way by machine learning algorithms. Section 5 outlines the application of machine learning for the data-driven detection of effect heterogeneities across subgroups defined upon observed covariates as well as for learning

optimal policy rules to target subgroups in a way that maximizes the treatment effect.

Section 6 considers treatment evaluation based on instrumental variables. Here, treatment selection may be related to unobserved characteristics if a quasi-random instrument exists that affects the treatment, but not directly the outcome. Section 7 discusses difference-in-differences methods, where identification hinges on common trends in mean potential outcomes under nontreatment over time across actually treated and nontreated groups. It also presents the changes-in-changes approach, which assumes that within treatment groups, the distribution of unobserved characteristics that affect the potential outcome under nontreatment remains constant over time. Section 8 introduces the regression discontinuity design, which assumes the treatment probability to discontinuously change and be quasi-randomly assigned at a specific threshold value of an observed index variable. It also discusses the regression kink design, which assumes a kink in the (continuous) association of the treatment and the index variable at a specific threshold. Section 9 concludes.

2 Selection on observables with preselected covariates

The selection-on-observables assumption, also called conditional independence or exogeneity, postulates that the covariate information in the data is rich enough to control for characteristics jointly affecting the treatment and the outcome. This implies that one either directly observes those characteristics confounding the treatment-outcome relationship or that conditional on the observed information, the effects of unobserved confounders on either the treatment or the outcome (or both) are blocked. As a further assumption known as common support, it is required that for any empirically feasible combination of observed covariates, both treated and nontreated subjects can be observed, which rules out that the covariates perfectly predict participation. Finally, the covariates must in general not be affected by the treatment, but measured at or prior to treatment assignment.

Denote by X the vector of observed covariates and $X(1), X(0)$ the potential covariate values

with and without treatment. Formally, the assumptions can be stated as

$$\{Y(1), Y(0)\} \perp D|X, \quad 0 < p(X) < 1, \quad X(1) = X(0) = X, \tag{5}$$

where $p(X) = \Pr(D = 1|X)$ is the conditional treatment probability, also known as propensity score. The first part of (5) means that the distributions of the potential outcomes are conditionally independent of the treatment. This implies that D is as good as randomly assigned among subjects with the same values in X. The second part says that the propensity score is larger than zero and smaller than one such that D is not deterministic in X. The third part states that X is not a function of D and therefore must not contain (post-treatment) characteristics that are affected by the treatment, in order to not condition away part of the treatment effect of interest. This identification approach mimics the experimental context with the help of observed information. After creating groups with and without treatment that are comparable in the covariates, differences in the outcomes are assumed to be exclusively caused by the treatment.

The first part of (5) is somewhat stronger than actually required for ATE identification and could be relaxed to conditional independence in the means (rather than all moments) of potential outcomes, $E[Y(d)|D = 1, X] = E[Y(d)|D = 0, X]$ for $d \in \{1, 0\}$. In empirical applications it might, however, be hard to argue that conditional independence holds in means but not in other distributional features, which would for instance rule out mean independence for nonlinear (e.g. log) transformations of Y. Furthermore, the stronger conditional independence assumption in (5) is required for the identification of distributional parameters like the quantile treatment effect, which corresponds to the effect at a particular rank of the potential outcome distribution. Also note that for the identification of treatment parameters among the treated (rather than the total) population like the ATET, (5) can be relaxed to $Y(1) \perp D|X$, $p(X) < 1$.

Let $\mu_d(x) = E[Y|D = d, X = x]$ denote the conditional mean outcome given D corresponding to $d \in \{1, 0\}$ and X equaling some value x in its support. Analogous to identification under a random treatment discussed in Section 1, $\mu_1(x) - \mu_0(x)$ under (5) identifies the conditional

average treatment effect (CATE) given X, denoted by Δ_x:

$$\Delta_x = E[Y(1) - Y(0)|X = x] = \mu_1(x) - \mu_0(x). \tag{6}$$

Averaging CATEs over X in the population or among treated yields the ATE or ATET, respectively:

$$\begin{aligned} \Delta &= E[\mu_1(X) - \mu_0(X)], \\ \Delta_{D=1} &= E[\mu_1(X) - \mu_0(X)|D = 1] = E[Y|D = 1] - E[\mu_0(X)|D = 1]. \end{aligned} \tag{7}$$

Noting that the propensity score possesses the so-called balancing property, see Rosenbaum & Rubin (1983), such that conditioning on $p(X)$ equalizes or balances the distribution of X across treatment groups (i.e. $X \perp D|p(X)$), the effects are also identified when substituting control variables X by $p(X)$:

$$\begin{aligned} \Delta &= E[\mu_1(p(X)) - \mu_0(p(X))], \\ \Delta_{D=1} &= E[\mu_1(p(X)) - \mu_0(p(X))|D = 1] = E[Y|D = 1] - E[\mu_0(p(X))|D = 1]. \end{aligned} \tag{8}$$

By basic probability theory, implying e.g. $\mu_1(X) = E[Y \cdot D|X]/p(X)$, and the law of iterated expectations, the ATE and ATET are also identified by inverse probability weighting (IPW), see Horvitz & Thompson (1952), using the propensity score:

$$\begin{aligned} \Delta &= E\left[\frac{Y \cdot D}{p(X)} - \frac{Y \cdot (1 - D)}{1 - p(X)}\right], \\ \Delta_{D=1} &= E\left[\frac{Y \cdot D}{\Pr(D = 1)} - \frac{Y \cdot (1 - D) \cdot p(X)}{(1 - p(X)) \cdot \Pr(D = 1)}\right]. \end{aligned} \tag{9}$$

Finally, the effects can be obtained from a combination of conditional mean outcomes and propensity scores related to the so-called efficient score function, see Robins, Rotnitzky & Zhao

(1994), Robins & Rotnitzky (1995), and Hahn (1998):

$$\begin{aligned}\Delta &= E\left[\phi(X)\right], \text{ with } \phi(X) = \mu_1(X) - \mu_0(X) + \frac{(Y-\mu_1(X))\cdot D}{p(X)} - \frac{(Y-\mu_0(X))\cdot(1-D)}{1-p(X)},\\ \Delta_{D=1} &= E\left[\frac{(Y-\mu_0(X))\cdot D}{\Pr(D=1)} - \frac{(Y-\mu_0(X))\cdot(1-D)\cdot p(X)}{(1-p(X))\cdot\Pr(D=1)}\right]. \qquad (10)\end{aligned}$$

Note that the identification results in (10) coincide with those in (9) and (7) because

$$\begin{aligned}&E\left[\frac{(Y-\mu_1(X))\cdot D}{p(X)} - \frac{(Y-\mu_0(X))\cdot(1-D)}{1-p(X)}\right] = 0 \quad \text{and}\\ &E\left[\frac{-\mu_0(X)\cdot D}{\Pr(D=1)} - \frac{-\mu_0(X)\cdot(1-D)\cdot p(X)}{(1-p(X))\cdot\Pr(D=1)}\right] = E\left[\mu_0(X)\cdot\left(\frac{p(X)}{\Pr(D=1)} - \frac{p(X)}{\Pr(D=1)}\right)\right] = 0.\end{aligned}$$

Assuming the availability of a randomly drawn sample, treatment effect estimation proceeds using the sample analogs of the identification results and plug-in estimates for $p(X), \mu_1(X), \mu_0(X)$ whenever required. When for instance considering the estimation of $\Delta_{D=1}$ based on (7), an estimate of $\mu_0(X)$ for each treated observation is obtained as a weighted average of nontreated outcomes, where the weights depend on the similarity of the treated and nontreated observations in terms of X. One class of methods in this context are matching estimators, see for instance Rosenbaum & Rubin (1983), Rosenbaum & Rubin (1985), Heckman, Ichimura & Todd (1998), Heckman, Ichimura, Smith & Todd (1998), Dehejia & Wahba (1999), and Lechner, Miquel & Wunsch (2011). Pair matching, for instance, assigns a weight of 1 (or 100%) to the most similar nontreated observation and of 0 to all others. $1:M$ matching estimates $\mu_0(X)$ based on the mean outcome of the M most similar nontreated observations, where M is an integer larger than 1. Radius or caliper matching defines a maximum tolerance of dissimilarity in X and relies on the mean outcome of all nontreated observations within the tolerance. Compared to $1:M$ estimation, this may reduce the variance when many similar nontreated observations are available. Due to the multidimensionality of X, similarity is to be defined by a distance metric. Examples include the square root of the sum of squared differences in elements of X across some treated and nontreated observation, either normalized by the inverse of the sample covariance matrix of

X (then called Mahalanobis distance) or by the diagonal thereof (i.e. the variance). See Zhao (2004) for a discussion of alternative distance metrics.

Abadie & Imbens (2006) show that in contrast to other treatment estimators, pair or $1:M$ matching generally does not converge with a rate of $n^{-1/2}$ to the true effect (i.e. is not $n^{-1/2}$-consistent) if X contains several continuous elements, with n being the sample size. Second, even under $n^{-1/2}$-consistency, it does not attain the semiparametric efficiency bounds derived in Hahn (1998). Therefore, pair or $1:M$ matching has a higher large sample variance than the most efficient (or least noisy) treatment effect estimators that rely on the same assumptions. Third, Abadie & Imbens (2008) demonstrate that bootstrapping, a popular inference method based on estimating the standard error based on repeatedly resampling from the data, is inconsistent due to the discontinuous weights in pair and $1:M$ matching. The authors, however, provide a consistent asymptotic approximation of the estimator's variance based on matching within treatment groups.

To improve upon its properties, matching can be combined with a regression-based correction of the bias that stems from not fully comparable treated and nontreated matches, see Rubin (1979) and Abadie & Imbens (2011). This matching-weighted regression is $n^{-1/2}$-consistent and its weights are smooth such that bootstrap inference is consistent. Another smooth method is kernel matching, which estimates $\mu_0(X)$ by a kernel function giving more weight to nontreated observations that are more similar to the treated reference observation and can attain the semiparametric efficiency bound. This requires no distance metric, as kernel functions are applied to each element in X and then multiplied. Finally, genetic matching of Diamond & Sekhon (2013) matches treated and nontreated observations in a way that maximizes the balance of covariate distributions across treatment groups according to predefined balance metrics, based on an appropriately weighted distance metric.

In empirical applications, matching on the estimated propensity score is much more common than matching directly on X. The propensity score is typically specified parametrically by logit or probit functions. Collapsing the covariate information into a single parametric function avoids

the curse of dimensionality, which implies that in finite samples, the probability of similar matches in all elements of X quickly decreases in the dimension of X. At the same time, it allows for effect heterogeneity across X. On the negative side, a misspecification of the propensity score model may entail an inconsistent treatment effect estimator, which is avoided by directly matching on X or using a nonparametric propensity score estimate. Matching on the estimated propensity score has a different variance than matching directly on X, which for the ATET can be either higher or lower, see Heckman, Ichimura & Todd (1998). Abadie & Imbens (2016) provide an asymptotic variance approximation for propensity score matching that appropriately accounts for uncertainty due to propensity score estimation.

Matching estimators typically require the choice of tuning parameters, be it the number of matches M, the bandwidth in kernel or radius matching, or the distance metric. However, theoretical guidance is frequently not available, see Frölich (2005) for an exception. Practitioners commonly pick tuning parameters ad hoc or based on data-driven methods that are not necessarily optimal for treatment effect estimation, as e.g. cross-validation for estimating $\mu_0(X)$. It appears thus advisable to investigate the sensitivity of the effect estimates w.r.t. varying these parameters.

As an alternative to matching, Hirano, Imbens & Ridder (2003) discuss treatment effect estimation based on the IPW sample analog of (9), using series regression to obtain nonparametric plug-in estimates of the propensity score, which attains the semiparametric efficiency bounds. Ichimura & Linton (2005) and Li, Racine & Wooldridge (2009) consider IPW with kernel-based propensity score estimation. Practitioners mostly rely on logit or probit specifications, which generally is not semiparametrically efficient, see Chen, Hong & Tarozzi (2008). In any case, it is common and recommended to use normalized sample analogs of the expressions in (9), which ensures that the weights of observations within treatment groups sum up to one, see Busso, DiNardo & McCrary (2014). Compared to matching, IPW has the advantages that it is computationally inexpensive and does not require choosing tuning parameters (other than for nonparametric propensity score estimation, if applied). On the negative side, IPW is likely sensitive to propensity scores that are very close to one or zero, see

the simulations in Frölich (2004) and Busso et al. (2014) and the theoretical discussion in Khan & Tamer (2010). Furthermore, IPW may be less robust to propensity score misspecification than matching, which merely uses the score to match treated and non-treated observations, rather than plugging it directly into the estimator, see Waernbaum (2012).

A variation of IPW are the empirical likelihood methods of Graham, Pinto & Egel (2012) and Imai & Ratkovic (2014). In spirit comparable to genetic matching, the methods iterate an initial propensity score estimate (e.g. by changing the coefficients of a logit specification) until prespecified moments of X are maximally balanced across treatment groups. A related approach is entropy balancing, see Hainmueller (2012), which iterates initially provided (e.g. uniform) weights until balance in the moments of X is maximized, under the constraint that weights sum up to one in either treatment group. In contrast to methods aiming for perfect covariate balance in prespecified moments, Zubizarreta (2015) trades off balance and variance in estimation. The algorithm finds the weights of minimum variance that balance the empirical covariate distribution up to prespecified levels, i.e. approximately rather than exactly.

Estimation based on the sample analog of (10) with plug-in estimates for $p(X), \mu_1(X), \mu_0(X)$ is called doubly robust (DR) estimation, as it is consistent if either the conditional mean outcome or the propensity score is correctly specified, see Robins, Mark & Newey (1992) and Robins, Rotnitzky & Zhao (1995). If both are correctly specified, DR is semiparametrically efficient. This is also the case if the plug-in estimates are nonparametrically estimated, see Cattaneo (2010). Furthermore, Rothe & Firpo (2013) show that nonparametric DR has a lower first order bias and second order variance than either IPW using a nonparametric propensity score or nonparametric outcome regression. This latter property is relevant in finite samples and implies that the accuracy of the DR estimator is less dependent on the accuracy of the plug-in estimates, e.g. the choice of the bandwidth in the kernel-based estimation of propensity scores and conditional mean outcomes. A further method satisfying the DR property is targeted maximum likelihood (TMLE), see van der Laan & Rubin (2006), in which an initial regression estimate is updated (or robustified) based on an IPW parameter.

3 Practical issues and extensions

This section discusses practical issues related to propensity score methods as well as extensions of treatment evaluation to non-binary treatments and different effect parameters. One important question is whether the estimated propensity score successfully balances X across treatment groups, e.g. in matched samples or after reweighting covariates (rather than outcomes) by IPW. Practitioners frequently consider hypothesis tests, e.g. two-sample t-tests applied to each element in X or F-tests for jointly testing imbalances in X, see also the joint tests of Sianesi (2004) and Smith & Todd (2005). As an alternative to hypothesis tests, Rosenbaum & Rubin (1985) consider a covariate's absolute mean difference across treated and nontreated matches, divided or standardized by the square root of half the sum of the covariate's variances in either treatment group prior to matching. In contrast to a t-test, which rejects balance under the slightest difference if the sample grows to infinity, this standardized difference is insensitive to the sample size. Rather than judging balance based on a p-value as in hypothesis tests, a standardized difference larger than a specific threshold, say 0.2, may be considered as indication for imbalance. On the negative side, the choice of the threshold appears rather arbitrary and data-driven methods for its determination are currently lacking. Taking the average of standardized differences for each covariate permits constructing a joint statistic for all covariates.

A second practical issue is whether common support in the propensity score distributions across treatment groups is sufficiently decent in the data. For the ATET, this implies that for each treated observation, nontreated matches with similar propensity scores exist, while for the ATE, this also needs to hold vice versa. Strictly speaking, common support is violated whenever for any reference observation, no observation in the other treatment group with exactly the same propensity score is available. In practice, propensity scores should be sufficiently similar, which requires defining a criterion based on which dissimilar observations may be discarded from the data to enforce common support. However, discarding observations implies that effect estimation might not be (fully) representative for the initial target population and thus sacrifices (some) external validity. On the other hand, it likely reduces estimation bias within the subpopulation satisfying

common support, thus enhancing internal validity. For possible common support criteria, see for instance Heckman, Ichimura, Smith & Todd (1998), who suggest discarding observations whose propensity scores have a density of or close to zero in (at least) one treatment group. For ATET estimation, Dehejia & Wahba (1999) propose discarding all treated observations with an estimated propensity score higher than the highest value among the nontreated. For the ATE, one additionally discards nontreated observations with a propensity score lower than the lowest value among the treated. Crump, Hotz, Imbens & Mitnik (2009) discuss dropping observations with propensity scores close to zero or one in a way that minimizes the variance of ATE estimation in the remaining sample. Huber, Lechner & Wunsch (2013) discard observations that receive a too large relative weight within their treatment group when estimating the treatment effect. See Lechner & Strittmatter (2019) for an overview of alternative common support criteria and an investigation of their performance in a simulation study.

The discussion so far focussed on a binary treatment, however, the framework straightforwardly extends to multivalued discrete treatments. The latter may either reflect distinct treatments (like different types of labor market programs as a job search training, a computer course, etc.) or discrete doses of a single treatment (like one, two, or three weeks of a training). Under appropriate selection-on-observable assumptions, treatment effects are identified by pairwise comparisons of each treatment value with nontreatment, or of two nonzero treatment values, if the effect of one treatment relative to the other is of interest. More formally, let d' and d'' denote the treatment levels to be compared and $I\{A\}$ the indicator function, which is one if event A holds and zero otherwise. Assume that conditions analogous to (5) are satisfied for $D = d'$ and $D = d''$, such that conditional independence assumptions $Y(d') \perp I\{D = d'\}|X$ and $Y(d'') \perp I\{D = d''\}|X$ hold and the so-called generalized propensity scores satisfy the common support restrictions $\Pr(D = d'|X) > 0$ and $\Pr(D = d''|X) > 0$, see Imbens (2000). Then, replacing D by $I\{D = d'\}$ and $1 - D$ by $I\{D = d''\}$ as well as $p(X) = \Pr(D = 1|X)$ by $\Pr(D = d'|X)$ and $1 - p(X)$ by $\Pr(D = d''|X)$ in the identification results (7), (8), (9), and (10) yields the ATE when comparing $D = d'$ vs. $D = d''$ as well as the ATET when considering those

with $D = d'$ as the treated. As shown in Cattaneo (2010), a range of treatment effect estimators for multivalued discrete treatments are $n^{-1/2}$-consistent and semiparametrically efficient under nonparametric estimation of the plug-in parameters. See also Lechner (2001) for a discussion of matching-based estimation with multivalued discrete treatments.

When D does not have discrete probability masses but is continuously distributed, the generalized propensity score corresponds to a conditional density, denoted by $f(D = d'|X)$ to distinguish it from the previously used probability $\Pr(D = d'|X)$. In the spirit of (7) for binary treatments, Flores (2007) proposes kernel regression of Y on D and X for estimating the mean potential outcomes of the continuous treatment. In analogy to (8), Hirano & Imbens (2005) regress Y on polynomials of D and estimates of $f(D|X)$ along with interactions, while Imai & van Dyk (2004) consider subclassification by the generalized propensity score. IPW-based methods as considered in Flores, Flores-Lagunes, Gonzalez & Neumann (2012) require replacing indicator functions, e.g. $I\{D = d'\}$, by continuous weighting functions in the identification results. Consider, for instance, the kernel weight $K\left((D - d')/h\right)/h$, where K is a symmetric second order kernel function (e.g. the standard normal density function) that assigns more weight to values of D the closer they are to d'. h is a bandwidth gauging by how quickly the weight decays as values in D become more different to d' and must go to zero as the sample size increases (albeit not too fast) for consistent estimation. Then, IPW-based identification of the ATE, for instance, corresponds to

$$\Delta = \lim_{h \to 0} E\left[\frac{Y \cdot K\left((D - d')/h\right)/h}{f(D = d'|X)} - \frac{Y \cdot K\left((D - d'')/h\right)/h}{f(D = d''|X))}\right], \tag{11}$$

where $\lim_{h \to 0}$ means 'as h goes to zero'. See Galvao & Wang (2015) for a further IPW approach and Kennedy, Ma, McHugh & Small (2017) for kernel-based DR estimation under continuous treatments, including data-driven bandwidth selection.

A further conceptual extension is the dynamic treatment framework, see for instance Robins (1986), Robins, Hernan & Brumback (2000), and Lechner (2009). It is concerned with

the evaluation of sequences of treatments (like consecutive labor market programs) based on sequential selection-on-observable assumptions w.r.t. each treatment. Related assumptions are also commonly imposed in causal mediation analysis aiming at disentangling a total treatment effect into various causal mechanisms, see for instance Robins & Greenland (1992), Pearl (2001), Imai, Keele & Yamamoto (2010), Tchetgen Tchetgen & Shpitser (2012), and Huber (2014), or the survey by Huber (2019). Finally, several contributions consider effect parameters related to distributions rather than means. Firpo (2007) proposes an efficient IPW estimator of quantile treatment effects (QTE) at specific ranks (like the median) of the potential outcome distribution and derives the semiparametric efficiency bounds. Donald & Hsu (2014) suggest IPW-based estimation of the distribution functions of potential outcomes under treatment and nontreatment, see also DiNardo, Fortin & Lemieux (1996) and Chernozhukov, Fernández-Val & Melly (2013) for estimators of counterfactual distributions. Imbens (2004) and Imbens & Wooldridge (2009) provide comprehensive reviews on treatment evaluation under selection on observables.

4 Causal machine learning

The treatment evaluation methods discussed so far consider covariates X as being preselected or fixed. This assumes away uncertainty related to model selection w.r.t. X and requires substantial or strictly speaking exact contextual knowledge about the confounders that need to be controlled for and in which functional form. In reality, however, practitioners frequently select covariates based on their predictive power for the treatment, typically without appropriately accounting for this model selection step in the causal inference to follow. Fortunately, this issue can be tackled by more recent treatment evaluation methods that incorporate machine learning to control for important confounders in a data-driven way and honestly account for model selection in the estimation process. This is particularly useful in big, and more specifically in wide data with a vast number of covariates that could potentially serve as control variables, which can render researcher-based covariate selection complicated if not infeasible.

It is important to see that when combining evaluation methods for the ATE or ATET with machine learning, henceforth called causal machine learning (CML), the data must contain sufficiently rich covariate information to satisfy the selection-on-observables assumption, just as discussed in Section 2. Therefore, CML is not a magic bullet that can do away with fundamental assumptions required for effect identification. However, it may be fruitfully applied if there exists a subset of covariate information that suffices to by and large tackle confounding, but is unknown to the researcher. Under the assumption that a relative to the sample size limited subset of information permits controlling for the most important confounders, CML can be shown to be approximately unbiased, even when confounding is not perfectly controlled for.

Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey & Robins (2018) consider for instance a CML approach called double machine learning that relies on so-called orthogonalized statistics. The latter imply that treatment effect estimation is rather insensitive to approximation errors in the estimation of $p(X), \mu_1(X), \mu_0(X)$. As discussed in Section 2, the sample analog of (10) satisfies this (doubly) robustness property along with its desirable finite sample behaviour. In contrast, estimation based on (7) is rather sensitive to approximation errors of $\mu_1(X), \mu_0(X)$, while estimation based on (9) is sensitive to errors in $p(X)$. Because DR, however, incorporates both propensity score and conditional mean outcome estimation, the approximation errors enter multiplicatively into the estimation problem, which is key for the robustness property, see for instance Farrell (2015).

A further element of many CML approaches including double machine learning is the use of independent samples for estimating the specifications of plug-in parameters like $p(X)$, $\mu_1(X)$, and $\mu_0(X)$ on the one hand and of the treatment effects $\Delta, \Delta_{D=1}$ on the other hand. This is similar in spirit to the idea of training and testing data in conventional machine learning or cross-validation for tuning parameter selection and obtained by randomly splitting the sample. After estimating models for $p(X), \mu_1(X), \mu_0(X)$ in one part of the data, the model parameters (e.g. coefficients) are used in the other part to predict $p(X), \mu_1(X), \mu_0(X)$ and ultimately estimate the treatment effect. Sample-splitting prevents overfitting the models for the plug-in parameters, but comes at

the cost that only part of the data are used for effect estimation, thus increasing the variance. So-called cross-fitting tackles this issue by swapping the roles of the data parts for estimating the plug-in models and the treatment effect. The treatment effect estimate is obtained as the average of the estimated treatment effects in each part and in fact, more than just two data splits may be used for this procedure. When combining DR with sample splitting, it suffices for $n^{-1/2}$-convergence of treatment effect estimation that the estimates of $p(X), \mu_1(X), \mu_0(X)$ converge to their respective true values at a rate of $n^{-1/4}$ (or faster), see Chernozhukov et al. (2018). Under specific regularity conditions, this convergence rate is attained by many machine learning algorithms and even by deep learning (which is popular in computer science e.g. for pattern recognition), see Farrell, Liang & Misra (2018).

However, it needs to be stressed that CML is conceptually different to standard machine learning, which aims at accurately predicting an outcome by observed predictors based on minimizing the prediction error (e.g. the mean squared error) through optimally trading off prediction bias and variance. This mere forecasting approach generally does not allow learning the causal effects of any of the predictors. One reason is that a specific predictor might obtain a smaller weight (e.g. regression coefficient) than implied by its true causal effect if the predictor is sufficiently correlated with other predictors, such that constraining its weight hardly affects the prediction bias, while reducing the variance. Therefore, predictive machine learning with Y as outcome and D and X as predictors generally gives a biased estimate of the causal effect of D, due to correlations between the treatment and the covariates. In CML, however, machine learning is not directly applied to ATE or ATET estimation, but merely for predicting the plug-in parameters, e.g. those of the DR expression (i.e. the sample analog of (10)) in the case of double machine learning. To this end, three separate machine learning predictions of D, Y among the treated, and Y among the nontreated are conducted with X being the predictors in each step. This is motivated by the fact that covariates X merely serve the purpose of tackling confounding, while their causal effects are (contrarily to the effect of D) not of interest, which makes the estimation of $p(X)$, $\mu_1(X)$, and $\mu_0(X)$ a prediction problem to which machine

learning can be applied.

Assume for instance that $\mu_1(X)$ and $\mu_0(X)$ are estimated by a linear lasso regression, see Tibshirani (1996), where X as well as higher order and interaction terms thereof may be included as predictors to allow for flexible model specifications. Including too many terms with low predictive power (as it would be the case in an overfitted polynomial regression) likely increases the variance of prediction, with little gain in terms of bias reduction. On the other hand, omitting important predictors implies a large increase in prediction bias relative to the gain in variance reduction due to a parsimonious specification. For this reason, lasso regression aims to optimally balance bias and variance through regularization, i.e. by shrinking the absolute coefficients obtained in a standard OLS regression towards or exactly to zero for less important predictors, e.g. based on cross-validation for determining the optimal amount of shrinkage. Analogously, lasso logit regression may be applied for the prediction of $p(X)$, which is a regularized version of a standard logit regression. Alternatively, lasso-based estimation of $\mu_1(X)$ and $\mu_0(X)$ can be combined with approximate covariate balancing of Zubizarreta (2015) instead of estimating a propensity score model for $p(X)$, see the CML algorithm suggested by Athey, Imbens & Wager (2018).

As discussed in Chernozhukov et al. (2018), lasso regression attains the required convergence rate of $n^{-1/4}$ under so-called approximate sparsity. The latter implies that the number of important covariates or interaction and higher order terms required for obtaining a sufficiently decent (albeit not perfect) approximation of the plug-in parameters is small relative to the sample size n. To see the merits of cross-fitting, note that when disregarding the latter and instead conducting the lasso and treatment estimation steps in the same (total) data, the number of important predictors is required to be small relative to $n^{-1/2}$ rather than n, see Belloni, Chernozhukov & Hansen (2014). Importantly, neither cross-fitting, nor the estimation of the plug-in parameters by some $n^{-1/4}$-consistent machine learning algorithm affects the asymptotic variance of treatment effect estimation (albeit it may matter in small samples). Therefore, CML is $n^{-1/2}$-consistent and attains the semiparametric efficiency bound as if the covariates to be controlled for in DR estimation had been correctly preselected. In large enough

samples, standard errors may thus be estimated by conventional asymptotic approximations without adjustment for the machine learning steps. For a more in depth review of various machine learning algorithms and CML, see for instance Athey & Imbens (2019).

5 Effect heterogeneity, conditional effects, and policy learning

Machine learning can also be fruitfully applied to investigate treatment effect heterogeneity across X, while possibly mitigating inferential multiple testing issues related to snooping for subgroups with significant(ly different) effects that might be spurious. For randomized experiments where (3) holds or under the selection-on-observables assumption (5) with preselected X, Athey & Imbens (2016) suggest a method that builds on a modification of so-called regression trees, see Breiman, Friedman, Olshen & Stone (1984). In standard machine learning for outcome prediction, the tree structure emerges by recursively partitioning the sample with respect to the predictor space such that the sum of squared deviations of outcomes and their respective partition means is minimized. This increases outcome homogeneity within and heterogeneity between partitions. Prediction of $E[Y|X = x]$ proceeds by taking the average of Y in the partition that includes the value $X = x$. This is equivalent to an OLS regression with predictors and interaction terms that are discretized according to specific threshold values in the covariate space as implied by the partitions. Cross-validation may be applied to find the optimal depth of partitions e.g. w.r.t. the mean squared error.

The causal tree approach of Athey & Imbens (2016) contains two key modifications when compared to standard regression trees. First, instead of Y, the mean difference in Y across treatment groups within partitions serves as outcome in the experimental context, while under selection on observables with preselected X, outcomes are reweighted by the inverse of the propensity score (in analogy to 9) prior to taking mean differences. In either case, recursive partitioning increases the homogeneity in estimated treatment effects within and its heterogeneity between partitions, in order to find the largest effect heterogeneities across

subgroups defined in terms of X. Secondly, applying sample splitting in order to use different data parts for estimating (a) the tree's model structure and (b) the treatment effects within partitions prevents spuriously large effect heterogeneities due to overfitting.

Wager & Athey (2018) and Athey, Tibshirani & Wager (2019) provide a further approach for investigating effect heterogeneity that is based on the related concept of random forests, see Breiman (2001), and also applies under selection on observables when control variables are not preselected but to be learnt from the data, see Section 4. Random forests consist of randomly drawing many subsamples from the original data and estimating trees in each subsample. Differently to standard trees, only a random subset of predictors (rather than all) is considered at each partitioning step, which safeguards against heavily correlated trees across subsamples. Predictions are obtained by averaging over the predictions of individual trees, which makes the random forest a smooth estimator and also reduces the variance when compared to discrete partitioning of a single tree. Forest-based predictions can therefore be represented by smooth weighting functions that bear some resemblance with kernel regression.

More concisely, the so-called generalized random forest of Athey et al. (2019) proceeds as follows. First, both Y and D are predicted as a function of X using random forests and leave-one-out cross-fitting. The latter implies that the outcome or treatment of each observation is predicted based on all observations in the data but its own, in order to prevent overfitting when conditioning on X. Second, the predictions are used for computing residuals of the outcomes and treatments, which is in the spirit of orthogonalized statistics as discussed in the context of DR in Section 4. Third, the effect of the residuals of D on the residuals of Y is predicted as a function of X by another random forest that averages over a large number of causal trees with residualized outcomes and treatments that use different parts of the respective subsamples for tree-modelling and treatment effect estimation. Bluntly speaking, this method combines the idea of sample splitting and orthogonalization to control for important confounders as discussed in Section 4 with the approach of Athey & Imbens (2016) for finding effect heterogeneity.

When comparing a single causal tree and a generalized random forest, an advantage of the

former is that it directly yields an easy-to-interpret partitioning based on the most predictive covariates in terms of effect heterogeneity. On the negative side, tree structures frequently have a rather high variance such that a small change in the data may entail quite different partitions. The generalized random forest is more attractive in terms of variance, but does not provide a single covariate partitioning due to averaging over many trees. It, however, yields an estimate of the CATE $\Delta_x = E[Y(1) - Y(0)|X = x]$, see (6), such that its heterogeneity as a function of X can be investigated. Also note that averaging over the estimates of Δ_x in the total sample or among the treatment provides consistent estimates of the ATE and ATET, respectively. For surveys on further machine learning methods for investigating treatment effect heterogeneity, see for instance Powers, Qian, Jung, Schuler, Shah, Hastie & Tibshirani (2018) and Knaus, Lechner & Strittmatter (2018).

A concept related to the CATE is optimal policy learning, see e.g. Manski (2004), Hirano & Porter (2009), Stoye (2009), Qian & Murphy (2011), Bhattacharya & Dupas (2012), and Kitagawa & Tetenov (2018), which typically aims at optimally allocating a costly treatment in some population under budget constraints. This for instance requires analyzing which observations in terms of covariate values X should be assigned the constrained treatment to maximize the average outcome. Examples include the optimal selection of jobseekers to be trained to maximize the overall employment probability or the optimal choice of customers to be offered a discount in order to maximize average sales. Formally, let $\pi'(X)$ denote a specific treatment policy defined as function of X. To give just one example, $\pi(X)$ could require $D = 1$ for all observations whose first covariate in X is larger than a particular threshold and $D = 0$ otherwise. The average effect of policy $\pi'(X)$, denoted by $Q(\pi'(X))$, corresponds to the difference in mean potential outcomes under $\pi(X)$ vs. nontreatment of everyone:

$$Q(\pi'(X)) = E[Y(\pi'(X)) - Y(0)] = E[\pi(X) \cdot \Delta_X]. \tag{12}$$

The second equality highlights the close relationship of policy learning and CATE identification. The optimal policy, denoted by $\pi^*(X)$, maximizes the average effect among the set of all feasible

policies contained in the set Π:

$$\pi^*(X) = \max_{\pi \in \Pi} Q(\pi(X)). \tag{13}$$

(12) and (13) permit defining the so-called regret function associated with treatment policy $\pi'(X)$, which is denoted by $R\pi'(X)$ and equals the (undesirable) reduction in the average policy effect due to implementing $\pi'(X)$ rather than the optimal policy $\pi^*(X)$:

$$R(\pi'(X)) = Q(\pi^*(X)) - Q(\pi'(X)). \tag{14}$$

Finding the optimal policy among the set of feasible policies Π, which implies that the average policy effect Q is maximized and regret R is equal to zero, amounts to solving the following maximization problem:

$$\pi^*(X) = \max_{\pi \in \Pi} E[(2\pi(X) - 1) \cdot \phi(X)]. \tag{15}$$

Note that $\phi(X)$ is the DR statistic of (10), see for instance Dudík, Langford & Li (2011), Zhang, Tsiatis, Davidian, Zhang & Laber (2012), and Zhou, Mayer-Hamblett, Khan & Kosorok (2017) for DR-based policy learning. The term $(2\pi(X) - 1)$ implies that the CATEs of treated and nontreated subjects enter positively and negatively into the expectation, respectively. Maximizing the expectation therefore requires optimally trading off treated and nontreated subjects in terms of their CATEs when choosing the treatment policy among all feasible policies. Estimation of the optimal policy may be based on the sample analog of (15), where $\phi(X)$ is estimated by cross-fitting and machine learning-based prediction of the plug-in parameters as outlined in Section 4. Athey & Wager (2018) demonstrate that similar to ATE estimation, basing policy learning on DR machine learning has desirable properties under specific conditions, even if the important elements in X driving confounding and/or effect heterogeneity are a priori unknown. The regret of the estimated optimal policy in the data when compared to the true optimal policy $\pi^*(X)$

decays at rate $n^{-1/2}$ under selection on observables if all plug-in parameters are estimated at rate $n^{-1/4}$. Zhou, Athey & Wager (2018) show how this result extends to policy learning for multivalued discrete treatments as also considered in Kallus (2017).

6 Instrumental variables

The selection-on-observables assumption imposed in the previous sections fails if selection into treatment is driven by unobserved factors that affect potential outcomes conditional on X. As an example, consider an experiment with imperfect compliance in which access to a training program is randomly assigned, but a subset of jobseekers that are offered the training does not comply and decides to not participate. If compliance behaviour is driven by unobserved factors (e.g. ability or motivation) that also affect the outcome (e.g. employment), endogeneity jeopardizes a causal analysis based on a naive comparison of treated and nontreated outcomes even when controlling for observed characteristics. However, if mere treatment assignment satisfies a so-called exclusion restriction such that it does not directly affect the outcome other than through actual treatment participation, it may serve as instrumental variable (IV), denoted by Z, to identify the treatment effect among those complying with the assignment. The intuition of IV-based identification is that the effect of Z of Y, which is identified by the randomization of the instrument, only operates through the effect of Z on D among compliers due to the exclusion restriction. Therefore, scaling (or dividing) the average effect of Z on Y by the average effect of Z on D yields the average effect of D on Y among compliers, see Imbens & Angrist (1994) and Angrist, Imbens & Rubin (1996).

However, in many applications it may not appear credible that IV assumptions like random assignment hold unconditionally, i.e. without controlling for observed covariates. This is commonly the case in observational data in which the instrument is typically not explicitly randomized like in an experiment. For instance, Card (1995) considers geographic proximity to college as IV for the likely endogenous treatment education when assessing its effect on earnings. While proximity might induce some individuals to go to college who would otherwise not, e.g. due to

housing costs associated with not living at home, it likely reflects selection into neighborhoods with a specific socio-economic status that affects labor market performance, implying that the IV is not random. If all confounders of the instrument-outcome relationship are plausibly observed in the data, IV-based estimation can be conducted conditional on observed covariates. For this reason, Card (1995) includes a range of control variables like parents' education, ethnicity, urbanity, and geographic region.

To formally state the IV assumptions that permit identifying causal effects conditional on covariates X in the binary instrument and treatment case, denote by $D(1)$ and $D(0)$ the potential treatment decision if instrument Z is set to 1 or 0, respectively. This permits defining four compliance types: Individuals satisfying $(D(1) = 1, D(0) = 0)$ are compliers as they only take the treatment when receiving the instrument. Non-compliers may consist of never takers who never take the treatment irrespective of the instrument $(D(1) = D(0) = 0)$, always takers $(D(1) = D(0) = 1)$, and defiers, who counteract instrument assignment $(D(1) = 0, D(0) = 1)$. Furthermore, denote (for the moment) the potential outcome as $Y(z, d)$, i.e. as function of both the instrument and the treatment. Then, the local average treatment effect (LATE) among compliers, denoted by $\Delta_{D(1)=1,D(0)=0} = E[Y(1) - Y(0)|D(1) = 1, D(0) = 0]$, is nonparametrically identified under the following assumptions, see Abadie (2003).

$$\begin{aligned} Z \perp (D(z), Y(z', d))|X \text{ for } z, z', d \in \{1, 0\}, \quad X(1) = X(0) = X, \\ \Pr(D(1) \geq D(0)|X) = 1, \quad E[D|Z = 1, X] - E[D|Z = 0, X] \neq 0, \\ \Pr(Y(1, d) = Y(0, d) = Y(d)|X) = 1 \text{ for } z, z', d \in \{1, 0\}. \end{aligned} \tag{16}$$

The first line of (16) says that conditional on X (which must not be affected by D), the IV is as good as random and thus not influenced by unobserved factors affecting the treatment and/or outcome. This is a selection-of-observables assumption similar to (5), however now imposed w.r.t. the instrument rather than the treatment. Therefore, the effects of Z on Y and on D are identified conditional on X, just in analogy to the identification of the effect of D on Y given X in Section

2. For this reason, replacing D by Z and the treatment propensity score $p(X) = \Pr(D = 1|X)$ by the instrument propensity score $\Pr(Z = 1|X)$ in the identification results for the ATE in (7), (8), (9), (10) yields the average effect of the instrument on the outcome. The latter is known as intention-to-treat effect (ITT) and henceforth denoted by θ. Additionally replacing Y by D yields the average effect of the instrument on the treatment (i.e. $E[D(1) - D(0)]$), the so-called first stage effect, denoted by γ.

The second line of (16) rules out the existence of defiers, but requires the existence of compliers conditional on X, due to the non-zero conditional first stage, while never and always takers might exist, too. By the law of total probability, this implies that γ corresponds to the share of compliers, as $D(1) - D(0)$ equals one for compliers and zero for never and always takers. The third line invokes the exclusion restriction such that Z must not have a direct effect on Y other than through D. By the law of total probability, the ITT in this case corresponds to the first stage effect γ times the LATE $\Delta_{D(1)=1,D(0)=0}$. This follows from the nonexistence of defiers and the fact that the effect of Z on Y is necessarily zero for always and never takers, whose D is not affected by Z. Therefore, the LATE is identified by scaling the ITT by the first stage effect. Formally,

$$\theta = \Delta_{D(1)=1,D(0)=0} \cdot \gamma \quad \Leftrightarrow \quad \Delta_{D(1)=1,D(0)=0} = \frac{\theta}{\gamma}. \tag{17}$$

If X is preselected, estimation of $\Delta_{D(1)=1,D(0)=0}$ proceeds by estimating both θ and γ based on any of the treatment effect estimators outlined in Section 2 and by dividing one by the other, which is $n^{-1/2}$-consistent under specific regularity conditions. Frölich (2007), for instance, considers nonparametric matching- and (local polynomial and series) regression-based estimation. Hong & Nekipelov (2010) derive semiparametric efficiency bounds for LATE estimation and propose efficient estimators. Donald, Hsu & Lieli (2014*b*) and Donald, Hsu & Lieli (2014*a*) propose IPW estimation using series logit and local polynomial regression-based estimation of the instrument propensity score. Tan (2006) and Uysal (2011) discuss DR estimation with parametric plug-in parameters. If IV confounders are not preselected but in analogy to Section 4 are to be

learnt from possibly high dimensional data, then causal machine learning may be applied to the DR representation of both θ and γ in order to estimate the LATE, see for instance Belloni, Chernozhukov, Fernández-Val & Hansen (2017). Finally, the analysis of effect heterogeneity and optimal policies discussed in Section 5 also extends to the IV context by using doubly robust statistics appropriate for LATE estimation, see Athey & Wager (2018) and Athey et al. (2019).

Frölich & Melly (2013) discuss the identification of the local quantile treatment effect on compliers (LQTE) and propose an IPW estimator based on local polynomial regression for IV propensity score estimation. Belloni et al. (2017) consider LQTE estimation based on causal machine learning when X are not preselected and important instrument confounders are to be learned from the data. In contrast to the previously mentioned studies, Abadie, Angrist & Imbens (2002) consider estimation of the conditional LQTE given particular values in X by applying the so-called κ-weighting approach of Abadie (2003). The latter permits identifying a broad class of complier-related statistics, based on the following weighting function κ:

$$\kappa = 1 - \frac{D \cdot (1 - Z)}{1 - \Pr(Z = 1|X)} - \frac{(1 - D) \cdot Z}{\Pr(Z = 1|X)}. \tag{18}$$

For instance, $\frac{E(\kappa \cdot X)}{E(\kappa)} = E[X|D(1) = 1, D(0) = 0]$ yields the mean of X among compliers, which permits judging the similarity of this subgroup and the total population in terms of observed characteristics.

The LATE assumptions are partly testable by investigating specific moment inequalities w.r.t. outcomes across complier types that need to hold for valid instruments, see the tests proposed by Kitagawa (2015), Huber & Mellace (2015), Mourifié & Wan (2017), Sharma (2016), and Guber (2018). The latter uses a modified version of the causal tree of Athey & Imbens (2016) to increase asymptotic power by searching for the largest violations in IV validity across values X in a data-driven way. It is also worth noting that even if monotonicity $\Pr(D(1) \geq D(0)|X) = 1$ is violated and defiers exist, the LATE on a fraction of compliers can still be identified if a subset of compliers is equal to the defiers in terms of the average effect and population size, see

de Chaisemartin (2017).

When extending the binary instrument and treatment case to a multivalued instrument Z and a binary D, LATEs are identified w.r.t. any pair of values (z'', z') satisfying the IV assumptions. Each of them may have a different first stage and thus, complier population. Particularly interesting appears the LATE for the largest possible complier population. The latter is obtained by defining the treatment propensity score $p(z) = \Pr(D = 1|Z = z, X = x)$ as instrument and considering the pair of propensity score values that maximizes compliance given $X = x$, see Frölich (2007).

A continuously distributed instrument even permits identifying a continuum of complier effects under appropriately adapted IV assumptions. Specifically, a marginal change in the instrument yields the so-called marginal treatment effect (MTE), see Heckman & Vytlacil (2001) and Heckman & Vytlacil (2005), which can be interpreted as the average effect among individuals who are indifferent between treatment or nontreatment given their values of Z and X. Technically speaking, the MTE is the limit of the LATE when the change in the instrument goes to zero.

In contrast to multivalued instruments, generalizing identification from binary to nonbinary treatments is not straightforward. Assume a binary instrument and an ordered treatment $D \in \{0, 1, ..., J\}$, with $J+1$ being the number of possible (discrete) treatment doses. Angrist & Imbens (1995) show that effects for single compliance types at specific treatment values, e.g. for those increasing the treatment from 1 to 2 when the increasing the instrument from 0 to 1, are not identified. It is, however, possible to obtain a non-trivially weighted average of effects of unit-level increases in the treatment on heterogeneous complier groups defined by different margins of the potential treatments. Albeit this is a proper causal parameter, its interpretability is compromised by the fact that the various complier groups generally enter with non-uniform weights. Similar issues occur if both instruments and treatments are multivalued.

There has been a controversial debate about the practical relevance of the LATE, as it only refers to the subgroup of compliers, see e.g. Deaton (2010), Imbens (2010), Heckman & Urzúa

(2010). It is therefore interesting to see under which conditions this effect can be extrapolated to other populations. As discussed in Angrist (2004), the LATE is directly externally valid, i.e., corresponds to the ATE when either all mean potential outcomes are homogeneous across compliance types, or at least the average effects. For testing the equality of mean potential outcomes across treated compliers and always takers as well as across nontreated compliers and never takers, see Angrist (2004), de Luna & Johansson (2014), Huber (2013), and Black, Joo, LaLonde, Smith & Taylor (2015). See also Donald et al. (2014*b*) for a related, but yet different testing approach. If equality in all mean potential outcomes holds at least conditional on X, instruments are in fact not required for identification as selection into D is on observables only, see Section 2. Angrist & Fernández-Val (2010) and Aronow & Carnegie (2013) do not consider homogeneity in mean potential outcomes but discuss extrapolation of the LATE when assuming homogeneous effects across compliance types. This assumption, which rules out selection into treatment by unobserved gains as assumed in standard Roy (1951) models, is testable if several instruments are available. For a comprehensive survey on methodological advancements in LATE evaluation, see Huber & Wüthrich (2019).

7 Difference-in-Differences

The difference-in-differences (DiD) approach bases identification on the so-called common trend assumption. The latter says that the mean potential outcomes under nontreatment of the actually treated and nontreated groups experience a common change over time when comparing periods before and after the treatment. Assuming that both groups would in the absence of the treatment have experienced the same time trend in potential outcomes, however, permits for differences in the levels of potential outcomes due to selection bias. As an example, assume that of interest is the employment effect of a minimum wage (D), which is introduced in one geographic region, but not in another one, see for instance Card & Krueger (1994). While the employment level (Y) may differ in both regions due to differences in the industry structure, DiD-based evaluation requires that employment changes e.g. due to business cycles would be the same in the absence

of a minimum wage. In this setup, a comparison of average employment in the post-treatment period across regions does not give the effect of the minimum wage due to selection bias related to the industry structure. A before-after comparison of employment (i.e. before and after treatment introduction) within the treated region is biased, too, as it picks up both the treatment effect and the business cycle-related time trend. Under the common trend assumption, however, the time trend for either region is identified by the before-after comparison in the nontreated region. Subtracting the before-after difference in employment in the nontreated region (time trend) from the before-after difference in the treated region (treatment effect plus time trend) therefore gives the treatment effect on the treated. That is, taking the difference in (before-after) differences across regions yields identification under the common trend assumption.

In many empirical problems, common trends may only appear plausible after controlling for observed covariates X. For instance, it could be argued that the assumption is more likely satisfied for treated and nontreated subjects within the same occupation or industry. Formally, let T denote a time index which is equal to zero in the pre-treatment period, when neither group received the treatment, and one in the post-treatment period, after one out of the two groups received the treatment. To distinguish the potential outcomes in terms of pre- and post-treatment periods, the subindex $t \in \{1, 0\}$ is added, such that $Y_0(1), Y_0(0)$ and $Y_1(1), Y_1(0)$ correspond to the pre- and post-treatment potential outcomes, respectively. The following conditions permit identifying the ATET in the post-treatment period, denoted by $\Delta_{D=1,T=1} = E[Y_1(1) - Y_1(0)|D = 1, T = 1]$, see the review of the DiD framework in Lechner (2010):

$$\begin{aligned} E[Y_1(0) - Y_0(0)|D = 1, X] &= E[Y_1(0) - Y_0(0)|D = 0, X], \quad X(1) = X(0) = X, \\ E[Y_0(0) - Y_0(0)|D = 1, X] &= 0, \\ \Pr(D = 1, T = 1|X, (D, T) &\in \{(d, t), (1, 1)\}) < 1 \text{ for all } (d, t) \in \{(1, 0), (0, 1), (0, 0)\}. \end{aligned} \tag{19}$$

The first line of (19) imposes that X is not affected by D and formalizes the conditional common trend assumption stating that conditional on X, no unobservables jointly affect the

treatment and the trend of mean potential outcomes under nontreatment. This is a selection-on-observables assumption on D, however, w.r.t. the changes in mean potential outcomes over time, rather than their levels as in (5) of Section 2. The two types of assumptions are not nested, such that neither implies the other, and cannot be combined for the sake of a more general model, see the discussion in Chabé-Ferret (2017). The second line in (19) rules out (average) anticipation effects among the treated, implying that D must not causally influence pre-treatment outcomes in expectation of the treatment to come. The third line imposes common support: For any value of X appearing in the group with $(D = 1, T = 1)$, subjects with such values of X must also exist in the remaining three groups with $(D = 1, T = 0)$, $(D = 0, T = 1)$, and $(D = 0, T = 0)$.

Given that the identifying assumptions hold, the DiD strategy applies to both panel data with the same subjects in pre- and post-treatment periods as well as to repeated cross sections with different subjects in either period. Under (19), $E[Y|D = 0, T = 1, X] - E[Y|D = 0, T = 0, X] = E[Y_1(0) - Y_0(0)|D = 0, X] = E[Y_1(0) - Y_0(0)|D = 1, X]$. This may be subtracted from $E[Y|D = 1, T = 1, X] - E[Y|D = 1, T = 0, X] = E[Y_1(1) - Y_0(1)|D = 1, X] = E[Y_1(1) - Y_1(0)|D = 1, X] + E[Y_1(0) - Y_0(1)|D = 1, X] = E[Y_1(1) - Y_0(1)|D = 1, X] = E[Y_1(1) - Y_1(0)|D = 1, X] + E[Y_1(0) - Y_0(0)|D = 1, X]$, where the second equality follows from subtracting and adding $Y_1(0)$ and the third from ruling out anticipation effects, in order to obtain the conditional ATET $E[Y_1(1) - Y_1(0)|D = 1, X]$. Therefore, averaging over the distribution of X among the treated in the post-treatment period yields the ATET in that period:

$$\begin{aligned} \Delta_{D=1,T=1} &= E[\mu_1(1, X) - \mu_1(0, X) - (\mu_0(1, X) - \mu_0(0, X))|D = 1, T = 1] \qquad (20) \\ &= E\left[\left\{\frac{D \cdot T}{\Pi} - \frac{D \cdot (1 - T) \cdot \rho_{1,1}(X)}{\rho_{1,0}(X) \cdot \Pi} - \left(\frac{(1 - D) \cdot T \cdot \rho_{1,1}(X)}{\rho_{0,1}(X) \cdot \Pi} - \frac{(1 - D) \cdot (1 - T) \cdot \rho_{1,1}(X)}{\rho_{0,0}(X) \cdot \Pi}\right)\right\} \cdot Y\right], \end{aligned}$$

where $\Pi = \Pr(D = 1, T = 1)$, $\rho_{d,t}(X) = \Pr(D = d, T = t|X)$, and $\mu_d(t, x) = E[Y|D = d, T = t, X = x]$.

As pointed out in Hong (2013), many DiD studies at least implicitly make the additional assumption that the joint distributions of treatment D and covariates X remain constant over

time T, formalized by $(X, D) \perp T$. This for instance rules out that the composition of X changes between periods in either treatment group. Under this additional assumption, $\Delta_{D=1,T=1}$ coincides with the 'standard' ATET $\Delta_{D=1}$, which is then identified by the following expressions:

$$\begin{aligned}
\Delta_{D=1} &= E[\mu_1(1,X) - \mu_1(0,X) - (\mu_0(1,X) - \mu_0(0,X))|D=1] \qquad (21)\\
&= E\left[\left\{\frac{D\cdot T}{P\cdot\Lambda} - \frac{D\cdot(1-T)}{P\cdot(1-\Lambda)} - \left(\frac{(1-D)\cdot T\cdot p(X)}{(1-p(X))\cdot P\cdot\Lambda} - \frac{(1-D)\cdot(1-T)\cdot p(X)}{(1-p(X))\cdot P\cdot(1-\Lambda)}\right)\right\}\cdot Y\right]\\
&= E\left[\left\{\frac{D\cdot T}{P\cdot\Lambda} - \frac{D\cdot(1-T)}{P\cdot(1-\Lambda)} - \left(\frac{(1-D)\cdot T\cdot p(X)}{(1-p(X))\cdot P\cdot\Lambda} - \frac{(1-D)\cdot(1-T)\cdot p(X)}{(1-p(X))\cdot P\cdot(1-\Lambda)}\right)\right\}\cdot(Y-\mu_0(T,X))\right],
\end{aligned}$$

where $p(X) = \Pr(D=1|X)$, $P = \Pr(D=1)$, and $\Lambda = \Pr(T=1)$. Exploiting the identification results after the first, second, and third equalities in (21), $n^{-1/2}$-consistent estimation may be based on regression or matching, on IPW as considered in Abadie (2005), or on DR estimation as in Sant'Anna & Zhao (2018), respectively. Zimmert (2018) shows that in the presence of high dimensional covariate information, causal machine learning based on the DR representation in (21) can be semiparametrically efficient in analogy to the results in Section 4.

A general practical issue concerning DiD inference is clustering, due to a correlation in uncertainty over time (e.g. in panel data due to having the same subjects in either period) or within regions (e.g. due to being exposed to the same institutional context). In this case, observations are not independently sampled from each other, implying that inference methods not accounting for clustering might perform poorly. See e.g. Bertrand, Duflo & Mullainathan (2004), Donald & Lang (2007), Cameron, Gelbach & Miller (2008), Conley & Taber (2011), Ferman & Pinto (2019) for a discussion of this issue as well as of (corrections of) asymptotic or bootstrap-based inference methods under a large or small number of clusters in the treatment groups. The findings of this literature suggest that cluster- and heteroskedasticity-robust variance estimators might only work satisfactorily if the number of treated and nontreated clusters is large enough, while a small number of clusters requires more sophisticated inference methods.

The subsequent discussion reviews some methodological extensions. de Chaisemartin & D'Haultfeuille (2018) discuss identification when the introduction of the treatment does not

induce everyone in the treatment group to be treated, but (only) increases the treatment rate more than in the nontreated group in the spirit of an instrument, see Section 6. Abraham & Sun (2018), Athey & Imbens (2018), Borusyak & Jaravel (2018), Callaway & Sant'Anna (2018), Goodman-Bacon (2018), Hull (2018), Strezhnev (2018), de Chaisemartin & D'Haultfeuille (2019), and Imai & Kim (2019) discuss DiD identification with multiple time periods and treatment groups that might experience treatment introduction at different points in time. Arkhangelsky, Athey, Hirshberg & Wager (2019) consider unit- and time-weighted DiD estimation.

Athey & Imbens (2006) suggest the so-called Changes-in-Changes (CiC) approach, which is related to DiD in that it exploits differences in pre- and post-treatment outcomes, however, based on different (and non-nested) identifying assumptions. While CiC does not invoke any common trend assumption, it imposes that potential outcomes under nontreatment are strictly monotonic in unobserved heterogeneity and that the distribution of the latter remains constant over time within treatment groups. Such a conditional independence between unobserved heterogeneity and time is satisfied if the subjects' ranks in the outcome distributions within treatment groups do not systematically change from pre- to post-treatment periods. In contrast to DiD, CiC allows identifying both the ATET and QTET, but generally requires a continuously distributed outcome for point identification.

Finally, another approach related to, but in terms of identification yet different from DiD is the synthetic control method of Abadie & Gardeazabal (2003) and Abadie, Diamond & Hainmueller (2010), which was originally developed for case study set ups with only one treated, but many nontreated units. It is based on appropriately weighting nontreated units to synthetically impute the treated unit's potential outcome under nontreatment. See e.g. the review article of Abadie & Cattaneo (2018) which contains a section on the synthetic control method that provides references to methodological advancements.

8 Regression discontinuity and kink designs

The regression discontinuity design (RDD), see Thistlethwaite & Campbell (1960), is based on the assumption that at a particular threshold of some observed running variable, the treatment status either changes from zero to one for everyone (sharp design) or for a subpopulation (fuzzy design). As an example, assume that the treatment of interest is extended eligibility to unemployment benefits, to which only individuals aged 50 or older are entitled, see for instance Lalive (2008). The idea is to compare the outcomes (like unemployment duration) of treated and untreated subjects close to the (age) threshold, e.g. of individuals aged 50 and 49, who are arguably similar in characteristics potentially affecting the outcome, due to their minor difference in age. The RDD therefore aims at imitating the experimental context at the threshold to evaluate the treatment effect locally for the subpopulation at the threshold.

Formally, let R denote the running variable and r_0 the threshold value. If the treatment is deterministic in R such that it is one whenever the threshold is reached or exceeded, i.e. $D = I\{R \geq r_0\}$, the RDD is sharp: All individuals change their treatment status exactly at r_0. Identification in the sharp RDD relies on the assumption that mean potential outcomes $E[Y(1)|R]$ and $E[Y(0)|R]$ are continuous and sufficiently smooth around $R = r_0$, see e.g. Hahn, Todd & van der Klaauw (2001), Porter (2003), and Lee (2008), meaning that any factors other than D that affect the outcome are continuous at the threshold. Continuity implies that if treated and nontreated populations with values of R exactly equal to r_0 existed, the treatment would be as good as randomly assigned w.r.t. mean potential outcomes. This corresponds to a local selection-on-observables assumption conditional on $R = r_0$. Furthermore, the density of the running variable R must be continuous and bounded away from zero around the threshold, such that treated and nontreated observations are observed close to $R = r_0$.

Under these assumptions, the ATE at the threshold, denoted by $\Delta_{R=r_0}$, is identified based on treated and nontreated outcomes in a neighbourhood $\varepsilon > 0$ around the threshold when letting ε

go to zero:

$$\begin{aligned} & \lim_{\varepsilon \to 0} E[Y|R \in [r_0, r_0 + \varepsilon)] - \lim_{\varepsilon \to 0} E[Y|R \in [r_0 - \varepsilon, r_0)] \qquad (22) \\ = \; & \lim_{\varepsilon \to 0} E[Y(1)|R \in [r_0, r_0 + \varepsilon)] - \lim_{\varepsilon \to 0} E[Y(0)|R \in [r_0 - \varepsilon, r_0)] = E[Y(1) - Y(0)|R = r_0] = \Delta_{R=r_0}. \end{aligned}$$

In the fuzzy RDD, D is not deterministic in R but may also depend on other factors. It is, however, assumed that the treatment share changes discontinuously at the threshold. Assume e.g. that admittance to a college (D) depends on passing a particular threshold of the score in a college entrance exam (R). While some students might decide not to attend college even if succeeding in the exam, a discontinuous change in the treatment share occurs if compliers exists that are induced to go to college when passing the threshold. Denote by $D(z)$ the potential treatment state as a function of the binary indicator $Z = I\{R \geq r_0\}$, which serves as instrument in an analogous way as discussed in Section 6. Similar to Dong (2014), assume that around the threshold, defiers do not exist and that the shares of compliers, always takers, and never takers as well as their mean potential outcomes under treatment and nontreatment are continuous. This implies that IV-type assumptions similar to those postulated in (16) conditional on X hold conditional on $R = r_0$.

Under these conditions, the first stage effect of Z on D, denoted by $\gamma_{R=r_0}$ is identified by

$$\begin{aligned} & \lim_{\varepsilon \to 0} E[D|R \in [r_0, r_0 + \varepsilon)] - \lim_{\varepsilon \to 0} E[D|R \in [r_0 - \varepsilon, r_0)] \qquad (23) \\ = \; & \lim_{\varepsilon \to 0} E[D(1)|R \in [r_0, r_0 + \varepsilon)] - \lim_{\varepsilon \to 0} E[D(0)|R \in [r_0 - \varepsilon, r_0)] = E[D(1) - D(0)|R = r_0] = \gamma_{R=r_0}. \end{aligned}$$

Furthermore, the first line of (22) identifies the ITT effect of Z on Y at the threshold, denoted by $\theta_{R=r_0}$ in the fuzzy RDD (rather than $\Delta_{R=r_0}$ as in the sharp RDD). In analogy to (17) in Section 6, the LATE on compliers at the treshold, denoted by $\Delta_{D(1)=1,D(0)=0,R=r_0} = E[Y(1) - Y(0)|D(1) = 1, D(0) = 0, R = r_0]$, is identified by dividing the ITT by the first stage effect at the threshold:

$$\Delta_{D(1)=1,D(0)=0,R=r_0} = \frac{\theta_{R=r_0}}{\gamma_{R=r_0}} \qquad (24)$$

In empirical applications of the RDD, the treatment effect is predominantly estimated by a local regression around the threshold. Practitioners for instance frequently use a linear regression for estimating $E[Y|D = 0, R < r_0]$ and $E[Y|D = 1, R \geq 0]$ within some bandwidth around r_0 in order to estimate $\Delta_{R=r_0}$ by the difference of the regression functions at r_0 in the case of the sharp RDD. A smaller bandwidth decreases estimation bias, because observations closer to the threshold are more comparable and effect estimation is more robust to model misspecification, see Gelman & Imbens (2018), but increases the variance due to relying on a lower number of observations. Imbens & Kalyanaraman (2012) propose a method for bandwidth selection that minimizes the squared error of the estimator. However, the optimal bandwidth for point estimation is generally suboptimal (and too large) for conducting inference, e.g. for computing confidence intervals. For this reason, Calonico, Cattaneo & Titiunik (2014) propose inference methods that are more robust to bandwidth choice and yield confidence intervals more closely matching nominal coverage, along with optimal bandwidth selection for inference. Their results imply that when $\Delta_{R=r_0}$ is estimated by linear regression within some bandwidth, then quadratic regression (i.e. one order higher) with the same bandwidth should be used for the computation of the standard error and confidence intervals. Armstrong & Kolesár (2018) suggest an alternative approach to inference that takes into account the worst case bias that could arise given a particular bandwidth choice. Cattaneo, Frandsen & Titiunik (2015) develop randomization methods for exact finite sample inference in the RDD under somewhat stronger identifying assumptions.

The identifying assumptions of the RDD are partly testable in the data. McCrary (2008) proposes a test for the continuity or the running variable at the threshold, as a discontinuity points to a manipulation of R and selective bunching at a one side of the threshold. In the previous example based on Lalive (2008), certain employees and companies might for instance manipulate age at entry into unemployment by postponing layoffs such that the age requirement for extended unemployment benefits is just satisfied. As a further test, Lee (2008) suggests investigating whether observed pre-treatment covariates X are locally balanced at either side of the threshold. Covariates also permit weakening the RDD assumptions to only hold conditional

on X, implying that all variables jointly affecting manipulation at the threshold and the outcome are observed, see Frölich & Huber (2018) who propose a nonparametric kernel estimator in this context. In contrast, Calonico, Cattaneo, Farrell & Titiunik (2018) do not exploit covariates for identification, but investigate variance reductions when linearly controlling for X and provide methods for optimal bandwidth selection and robust inference for this case.

Several studies investigate conditions under which the rather local RDD effect can be extrapolated to other populations. Dong & Lewbel (2015) show the identification of the derivative of the RDD treatment effect in both sharp and fuzzy designs, which permits identifying the change in the treatment effect resulting from a marginal change in the threshold. Angrist & Rokkanen (2015) test whether the running variable's association with the outcome vanishes on either side of the threshold conditional on covariates X. For the case of the sharp RDD, this implies that X is sufficient to control for confounding just as under the selection-on-observables framework of Section 2, such that effects are also identified away from the threshold. In context of the fuzzy RDD, Bertanha & Imbens (n.d.) propose a test for the equality in mean outcomes of treated compliers and always takers, as well as of untreated compliers and never takers. This permits investigating whether the effect on compliers at the threshold may be extrapolated to all compliance types at and away from the threshold. Cattaneo, Keele, Titiunik & Vazquez-Bare (2019) demonstrate extrapolation under multiple thresholds, i.e. when the threshold may vary for various subjects instead of being equal for everyone, as considered in Cattaneo, Keele, Titiunik & Vazquez-Bare (2016).

Lee & Card (2008), Dong (2015), Kolesár & Rothe (2018) discuss identification and inference when the forcing variable is discrete rather than continuous, which is highly relevant for empirical applications. Papay, Willett & Murnane (2011) and Keele & Titiunik (2015) extend the regression-discontinuity approach to multiple running variables. Imbens & Wager (2019) propose an optimization-based inference method for deriving the minimax linear RDD estimator which can be applied to continuous, discrete, and multiple running variables. Frandsen, Frölich & Melly (2012) discuss the identification of quantile treatment effects in the RDD. See also

Imbens & Lemieux (2008) and Lee & Lemieux (2010) for surveys on the applied and theoretical RDD literature.

Related to the fuzzy RDD is the regression kink design (RKD), see Card, Lee, Pei & Weber (2015), which is technically speaking a first derivative version of the former. The treatment is assumed to be a continuous function of the running variable R (rather than discontinuous as in the RDD), with a kink at r_0. This implies that the first derivative of D w.r.t. R (rather than the level of D as in the RDD) is discontinuous at the threshold. In Landais (2015), for instance, unemployment benefits (D) are a kinked function of the previous wage (R): D corresponds to R times a constant percentage up to a maximum previous wage r_0 beyond which D does not increase any further but remains constant. For this piecewise linear function, the derivative of D w.r.t. R corresponds to the percentage for $R < r_0$ and to zero for $R \geq 0$. As the treatment is deterministic in the running variable, this is known as sharp RKD.

Given appropriate continuity and smoothness conditions w.r.t. mean potential outcomes and the density of R around r_0, scaling the change in the first derivatives of mean outcomes w.r.t. to R at the threshold by the corresponding change in first derivatives of D identifies a causal effect. The latter corresponds to the average derivative of the potential outcome with respect to D when the latter corresponds to its value at the threshold, denoted by d_0, within the local population at $R = r_0$:

$$\Delta_{R=r_0}(d_0) = \frac{\partial E[Y(d_0)|R=r_0]}{\partial D} = \frac{\lim\limits_{\varepsilon\to 0}\frac{\partial E[Y|R\in[r_0,r_0+\varepsilon)]}{\partial R} - \lim\limits_{\varepsilon\to 0}\frac{\partial E[Y|R\in[r_0-\varepsilon,r_0)]}{\partial R}}{\lim\limits_{\varepsilon\to 0}\frac{\partial D|R\in[r_0,r_0+\varepsilon)}{\partial R} - \lim\limits_{\varepsilon\to 0}\frac{\partial D|R\in[r_0-\varepsilon,r_0)}{\partial R}} \tag{25}$$

The fuzzy RKD permits deviations from the kinked function characterizing how the running variable affects the treatment, such that D is not deterministic in R, see for instance Simonsen, Skipper & Skipper (2016) for a study investigating the price sensitivity of product demand. Under specific continuity conditions and the monotonicity-type assumption that the kink of any individual either goes in the same direction or is zero, a causal effect at the threshold is identified among individuals with nonzero kinks. To this end, the derivatives of the treatment in (25),

namely $\frac{\partial D|R\in[r_0,r_0+\varepsilon)}{\partial R}$ and $\frac{\partial D|R\in[r_0-\varepsilon,r_0)}{\partial R}$, are to be replaced by the derivatives of expectations $\frac{\partial E[D|R\in[r_0,r_0+\varepsilon)]}{\partial R}$ and $\frac{\partial E[D|R\in[r_0-\varepsilon,r_0)]}{\partial R}$. As the expectation of a treatment maybe continuous even if the treatment itself is not, the fuzzy RKD may also be applied to a binary D, see Dong (2014). Calonico et al. (2014) provide robust inference methods for the RKD, while Ganong & Jäger (2018) propose a permutation method for exact finite sample inference.

9 Conclusion

This chapter provided an overview of different approaches to policy evaluation for assessing the causal effect of a treatment on an outcome. Starting with an introduction to causality and the experimental evaluation of a randomized treatment, it subsequently discussed identification and flexible estimation under selection on observables, instrumental variables, difference-in-differences, changes-in-changes, and regression discontinuities and kinks. Particular attention was devoted to approaches combining policy evaluation with machine learning to provide data-driven procedures for tackling confounding related to observed covariates, investigating effect heterogeneities across subgroups, and learning optimal treatment policies. In a world with ever increasing data availability, such causal machine learning methods aimed at optimally exploiting large amounts of information for causal inference will likely leverage the scope of policy evaluation to unprecedented levels. Besides the classic domain of public policies, this concerns not least the private sector, with ever more firms investing in data analytics to assess and optimize the causal impact of their actions like price policies or advertising campaigns.

References

Abadie, A. (2003), 'Semiparametric instrumental variable estimation of treatment response models', *Journal of Econometrics* **113**, 231–263.

Abadie, A. (2005), 'Semiparametric difference-in-differences estimators', *Review of Economic Studies* **72**, 1–19.

Abadie, A., Angrist, J. & Imbens, G. W. (2002), 'Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings', *Econometrica* **70**, 91 – 117.

Abadie, A. & Cattaneo, M. D. (2018), 'Econometric methods for program evaluation', *Annual Review of Economics* **10**, 465–503.

Abadie, A., Diamond, A. & Hainmueller, J. (2010), 'Synthetic control methods for comparative case studies: Estimating the effect of californias tobacco control program', *Journal of the American Statistical Association* **105**, 493–505.

Abadie, A. & Gardeazabal, J. (2003), 'The economic costs of conflict: A case study of the basque country', *The American Economic Review* **93**, 113–132.

Abadie, A. & Imbens, G. W. (2006), 'Large sample properties of matching estimators for average treatment effects', *Econometrica* **74**, 235–267.

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

Abadie, A. & Imbens, G. W. (2011), 'Bias-corrected matching estimators for average treatment effects', *Journal of Business & Economic Statistics* **29**, 1–11.

Abadie, A. & Imbens, G. W. (2016), 'Matching on the estimated propensity score', *Econometrica* **84**, 781–807.

Abraham, S. & Sun, L. (2018), 'Estimating dynamic treatment effects in event studies with heterogeneous treatment effects', *working paper, Massachusetts Institute of Technology* .

Angrist, J. D. (2004), 'Treatment effect heterogeneity in theory and practice', *The Economic Journal* **114**, C52–C83.

Angrist, J. D. & Rokkanen, M. (2015), 'Wanna get away? regression discontinuity estimation of exam school effects away from the cutoff', *Journal of the American Statistical Association* **110**, 1331–1344.

Angrist, J. & Fernández-Val, I. (2010), 'Extrapolate-ing: External validity and overidentification in the late framework', *NBER working paper 16566* .

Angrist, J., Imbens, G. & Rubin, D. (1996), 'Identification of causal effects using instrumental variables', *Journal of American Statistical Association* **91**, 444–472 (with discussion).

Angrist, J. & Imbens, G. W. (1995), 'Two-stage least squares estimation of average causal effects in models with variable treatment intensity', *Journal of American Statistical Association* **90**, 431–442.

Arkhangelsky, D., Athey, S., Hirshberg, D. A. & Wager, G. W. I. S. (2019), 'Synthetic difference in differences', *working paper, Stanford University* .

Armstrong, T. B. & Kolesár, M. (2018), 'Optimal inference in a class of regression models', *Econometrica* **86**(2), 655–683.

Aronow, P. M. & Carnegie, A. (2013), 'Beyond late: Estimation of the average treatment effect with an instrumental variable', *Political Analysis* **21**, 492–506.

Athey, S. & Imbens, G. (2006), 'Identification and inference in nonlinear difference-in-differences models', *Econometrica* **74**, 431–497.

Athey, S. & Imbens, G. (2016), 'Recursive partitioning for heterogeneous causal effects', *Proceedings of the National Academy of Sciences* **113**, 7353–7360.

Athey, S. & Imbens, G. (2018), 'Design-based analysis in difference-in-differences settings with staggered adoption', *working paper, Stanford University* .

Athey, S. & Imbens, G. W. (2019), 'Machine learning methods that economists should know about', *Annual Review of Economics* **11**.

Athey, S., Imbens, G. W. & Wager, S. (2018), 'Approximate residual balancing: debiased inference of average treatment effects in high dimensions', *Journal of the Royal Statistical Society Series B* **80**, 597–623.

Athey, S., Tibshirani, J. & Wager, S. (2019), 'Generalized random forests', *The Annals of Statistics* **47**, 1148–1178.

Athey, S. & Wager, S. (2018), 'Efficient policy learning', *working paper, Stanford University* .

Belloni, A., Chernozhukov, V., Fernández-Val, I. & Hansen, C. (2017), 'Program evaluation and causal inference with high-dimensional data', *Econometrica* **85**, 233–298.

Belloni, A., Chernozhukov, V. & Hansen, C. (2014), 'Inference on treatment effects after selection among high-dimensional controls', *The Review of Economic Studies* **81**, 608–650.

Bertanha, M. & Imbens, G. W. (n.d.), 'External validity in fuzzy regression discontinuity designs', *forthcoming in the Journal of Business & Economic Statistics* .

Bertrand, M., Duflo, E. & Mullainathan, S. (2004), 'How much should we trust differences-in-differences estimates?', *Quarterly Journal of Economics* **119**, 249–275.

Bhattacharya, D. & Dupas, P. (2012), 'Inferring welfare maximizing treatment assignment under budget constraints', *Journal of Econometrics* **167**, 168–196.

Black, D. A., Joo, J., LaLonde, R. J., Smith, J. A. & Taylor, E. J. (2015), 'Simple tests for selection bias: Learning more from instrumental variables', *IZA Discussion Paper No 9346* .

Borusyak, K. & Jaravel, X. (2018), 'Revisiting event study designs', *working paper, Harvard University* .

Breiman, L. (2001), 'Random forests', *Machine Learning* **45**, 5–32.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984), *Classification and Regression Trees*, Wadsworth, Belmont, California.

Busso, M., DiNardo, J. & McCrary, J. (2014), 'New evidence on the finite sample properties of propensity score matching and reweighting estimators', *Review of Economics and Statistics* **96**, 885–897.

Callaway, B. & Sant'Anna, P. H. C. (2018), 'Difference-in-differences with multiple time periods and an application on the minimum wage and employment', *working paper, Vanderbilt University* .

Calonico, S., Cattaneo, M. D., Farrell, M. H. & Titiunik, R. (2018), 'Regression discontinuity designs using covariates', *forthcoming in the Review of Economics and Statistics* .

Calonico, S., Cattaneo, M. D. & Titiunik, R. (2014), 'Robust nonparametric confidence intervals for regression-discontinuity designs', *Econometrica* **82**, 2295–2326.

Cameron, A. C., Gelbach, J. B. & Miller, D. L. (2008), 'Bootstrap-based improvements for inference with clustered errors', *Review of Economics and Statistics* **90**, 414–427.

Card, D. (1995), Using geographic variation in college proximity to estimate the return to schooling, *in* L. Christofides, E. Grant & R. Swidinsky, eds, 'Aspects of Labor Market Behaviour: Essays in Honour of John Vanderkamp', University of Toronto Press, Toronto, pp. 201–222.

Card, D. & Krueger, A. B. (1994), 'Minimum wages and employment: A case study of the fast-food industry in new jersey and pennsylvania', *The American Economic Review* **84**, 772–793.

Card, D., Lee, D. S., Pei, Z. & Weber, A. (2015), 'Inference on causal effects in a generalized regression kink design', *Econometrica* **83**, 2453–2483.

Cattaneo, M. D. (2010), 'Efficient semiparametric estimation of multi-valued treatment effects under ignorability', *Journal of Econometrics* **155**, 138 – 154.

Cattaneo, M. D., Frandsen, B. R. & Titiunik, R. (2015), 'Randomization inference in the regression discontinuity design: An application to party advantages in the u.s. senate', *Journal of Causal Inference* **3**.

Cattaneo, M. D., Keele, L., Titiunik, R. & Vazquez-Bare, G. (2016), 'Interpreting regression discontinuity designs with multiple cutoffs', *The Journal of Politics* **78**, 12291248.

Cattaneo, M. D., Keele, L., Titiunik, R. & Vazquez-Bare, G. (2019), 'Extrapolating treatment effects in multi-cutoff regression discontinuity designs', *working paper, University of Michigan* .

Chabé-Ferret, S. (2017), 'Should we combine difference in differences with conditioning on pre-treatment outcomes', *working paper, Toulouse School of Economics* .

Chen, X., Hong, H. & Tarozzi, A. (2008), 'Semiparametric efficiency in gmm models with auxiliary data', *The Annals of Statistics* **36**, 808–843.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. & Robins, J. (2018), 'Double/debiased machine learning for treatment and structural parameters', *The Econometrics Journal* **21**, C1–C68.

Chernozhukov, V., Fernández-Val, I. & Melly, B. (2013), 'Inference on counterfactual distributions', *Econometrica* **81**, 2205–2268.

Conley, T. & Taber, C. (2011), 'Inference with "difference in differences" with a small number of policy changes', *Review of Economics and Statistics* **93**, 113–125.

Crump, R., Hotz, J., Imbens, G. & Mitnik, O. (2009), 'Dealing with limited overlap in estimation of average treatment effects', *Biometrika* **96**, 187–199.

de Chaisemartin, C. (2017), 'Tolerating defiance? local average treatment effects without monotonicity', *Quantitative Economics* **8**, 367–396.

de Chaisemartin, C. & D'Haultfeuille, X. (2018), 'Fuzzy differences-in-differences', *Review of Economic Studies* **85**, 999–1028.

de Chaisemartin, C. & D'Haultfeuille, X. (2019), 'Two-way fixed effects estimators with heterogeneoustreatment effects', *working paper ,University of California at Santa Barbara* .

de Luna, X. & Johansson, P. (2014), 'Testing for the unconfoundedness assumption using an instrumental assumption', *Journal of Causal Inference* **2**, 187–199.

Deaton, A. S. (2010), 'Instruments, randomization, and learning about development', *Journal of Economic Literature* **48**, 424–455.

Dehejia, R. H. & Wahba, S. (1999), 'Causal effects in non-experimental studies: Reevaluating the evaluation of training programmes', *Journal of American Statistical Association* **94**, 1053–1062.

Diamond, A. & Sekhon, J. S. (2013), 'Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies', *Review of Economics and Statistics* **95**, 932–945.

DiNardo, J. E., Fortin, N. M. & Lemieux, T. (1996), 'Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach', *Econometrica* **64**, 1001–1044.

Donald, S. G. & Hsu, Y. C. (2014), 'Estimation and inference for distribution functions and quantile functions in treatment effect models', *Journal of Econometrics* **178**, 383–397.

Donald, S. G., Hsu, Y.-C. & Lieli, R. P. (2014*a*), 'Inverse probability weighted estimation of local average treatment effects: A higher order mse expansion', *Statistics and Probability Letters* **95**, 132–138.

Donald, S. G., Hsu, Y.-C. & Lieli, R. P. (2014*b*), 'Testing the unconfoundedness assumption via inverse probability weighted estimators of (L)ATT', *Journal of Business & Economic Statistics* **32**, 395–415.

Donald, S. & Lang, K. (2007), 'Inference with difference-in-differences and other panel data', *Review of Economics and Statistics* **89**, 221–233.

Dong, Y. (2014), 'Jumpy or kinky? regression discontinuity without the discontinuity', *working Paper, University of Califronia Irvine* .

Dong, Y. (2015), 'Regression discontinuity applications with rounding errors in the running variable', *Journal of Applied Econometrics* **30**, 422–446.

Dong, Y. & Lewbel, A. (2015), 'Identifying the effect of changing the policy threshold in regression discontinuity models', *Review of Economics and Statistics* **97**, 1081–1092.

Dudík, M., Langford, J. & Li, L. (2011), 'Doubly robust policy evaluation and learning', *Procceedings of the 28th International Conference on Machine Learning* pp. 1097–1104.

Farrell, M. H. (2015), 'Robust inference on average treatment effects with possibly more covariates than observations', *Journal of Econometrics* **189**, 1–23.

Farrell, M. H., Liang, T. & Misra, S. (2018), 'Deep neural networks for estimation and inference: Application to causal effects and other semiparametric estimands', *working paper, University of Chicago* .

Ferman, B. & Pinto, C. (2019), 'Inference in differences-in-differences with few treated groups and heteroskedasticity', *The Review of Economics and Statistics* **101**, 116.

Firpo, S. (2007), 'Efficient Semiparametric Estimation of Quantile Treatment Effects', *Econometrica* **75**, 259–276.

Flores, C. A. (2007), 'Estimation of dose-response functions and optimal doses with a continuous treatment', *working paper, University of California, Berkeley.* .

Flores, C. A., Flores-Lagunes, A., Gonzalez, A. & Neumann, T. C. (2012), 'Estimating the effects of length of exposure to instruction in a training program: the case of job corps', *The Review of Economics and Statistics* **94**, 153–171.

Frandsen, B. R., Frölich, M. & Melly, B. (2012), 'Quantile treatment effects in the regression discontinuity design', *Journal of Econometrics* **168**.

Frölich, M. (2004), 'Finite sample properties of propensity-score matching and weighting estimators', *The Review of Economics and Statistics* **86**, 77–90.

Frölich, M. (2005), 'Matching estimators and optimal bandwidth choice', *Statistics and Computing* **15**, 197–215.

Frölich, M. (2007), 'Nonparametric iv estimation of local average treatment effects with covariates', *Journal of Econometrics* **139**, 35–75.

Frölich, M. & Huber, M. (2018), 'Including covariates in the regression discontinuity design', *Journal of Business and Economic Statistics* .

Frölich, M. & Melly, B. (2013), 'Unconditional quantile treatment effects under endogeneity', *Journal of Business & Economic Statistics* **31**, 346–357.

Galvao, A. F. & Wang, L. (2015), 'Uniformly semiparametric efficient estimation of treatment effects with a continuous treatment', *Journal of the American Statistical Association* **110**, 1528–1542.

Ganong, P. & Jäger, S. (2018), 'A permutation test for the regression kink design', *Journal of the American Statistical Association* **113**, 494–504.

Gelman, A. & Imbens, G. (2018), 'Why high-order polynomials should not be used in regression discontinuity designs', *forthcoming in the Journal of Business & Economic Statistics* .

Goodman-Bacon, A. (2018), 'Difference-in-differences with variation in treatment timing', *working paper, Vanderbilt University* .

Graham, B., Pinto, C. & Egel, D. (2012), 'Inverse probability tilting for moment condition models with missing data', *Review of Economic Studies* **79**, 1053–1079.

Guber, R. (2018), 'Instrument validity tests with causal trees:with an application to the same-sex instrument', *working paper, Munich Center for the Economics of Aging* .

Hahn, J. (1998), 'On the role of the propensity score in efficient semiparametric estimation of average treatment effects', *Econometrica* **66**, 315–331.

Hahn, J., Todd, P. & van der Klaauw, W. (2001), 'Identification and estimation of treatment effects with a regression-discontinuity design', *Econometrica* **69**, 201–209.

Hainmueller, J. (2012), 'Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies', *Political Analysis* **20**, 25–46.

Heckman, J., Ichimura, H., Smith, J. & Todd, P. (1998), 'Characterizing selection bias using experimental data', *Econometrica* **66**, 1017–1098.

Heckman, J. J., Ichimura, H. & Todd, P. (1998), 'Matching as an econometric evaluation estimator', *Review of Economic Studies* **65**, 261–294.

Heckman, J. J. & Urzúa, S. (2010), 'Comparing iv with structural models: What simple iv can and cannot identify', *Journal of Econometrics* **156**, 27–37.

Heckman, J. J. & Vytlacil, E. (2001), Local instrumental variables, *in* C. Hsiao, K. Morimune & J. Powell, eds, 'Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya', Cambridge University Press, Cambridge.

Heckman, J. J. & Vytlacil, E. (2005), 'Structural equations, treatment effects, and econometric policy evaluation 1', *Econometrica* **73**, 669–738.

Hirano, K. & Imbens, G. W. (2005), *The Propensity Score with Continuous Treatments*, Wiley-Blackwell, chapter 7, pp. 73–84.

Hirano, K., Imbens, G. W. & Ridder, G. (2003), 'Efficient estimation of average treatment effects using the estimated propensity score', *Econometrica* **71**, 1161–1189.

Hirano, K. & Porter, J. (2009), 'Asymptotics for statistical treatment rules', *Econometrica* **77**, 1683–1701.

Hong, H. & Nekipelov, D. (2010), 'Semiparametric efficiency in nonlinear late models', *Quantitative Economics* **1**, 279–304.

Hong, S.-H. (2013), 'Measuring the effect of napster on recorded music sales: difference?in?differences estimates under compositional changes', *Journal of Applied Econometrics* **28**, 297–324.

Horvitz, D. & Thompson, D. (1952), 'A generalization of sampling without replacement from a finite population', *Journal of American Statistical Association* **47**, 663–685.

Huber, M. (2013), 'A simple test for the ignorability of non-compliance in experiments', *Economics Letters* **120**, 389–391.

Huber, M. (2014), 'Identifying causal mechanisms (primarily) based on inverse probability weighting', *Journal of Applied Econometrics* **29**, 920–943.

Huber, M. (2019), 'A review of causal mediation analysis for assessing direct and indirect treatment effects', *SES working paper 500, University of Fribourg* .

Huber, M., Lechner, M. & Wunsch, C. (2013), 'The performance of estimators based on the propensity score', *Journal of Econometrics* **175**, 1–21.

Huber, M. & Mellace, G. (2015), 'Testing instrument validity for late identification based on inequality moment constraints', *Review of Economics and Statistics* **97**, 398–411.

Huber, M. & Wüthrich, K. (2019), 'Local average and quantile treatment effects under endogeneity: A review', *Journal of Econometric Methods* **8**, 1–28.

Hull, P. (2018), 'Estimating treatment effects in mover designs', *working paper, University of Chicago* .

Ichimura, H. & Linton, O. (2005), Asymptotic expansions for some semiparametric program evaluation estimators, *in* D. Andrews & J. Stock, eds, 'Identification and Inference for Econometric Models', Cambridge University Press, Cambridge, pp. 149–170.

Imai, K., Keele, L. & Yamamoto, T. (2010), 'Identification, inference and sensitivity analysis for causal mediation effects', *Statistical Science* **25**, 51–71.

Imai, K. & Kim, I. S. (2019), 'On the use of two-way fixed effects regression models for causal inference with panel data', *working paper, Harvard University* .

Imai, K. & Ratkovic, M. (2014), 'Covariate balancing propensity score', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**, 243–263.

Imai, K. & van Dyk, D. A. (2004), 'Causal inference with general treatment regimes', *Journal of the American Statistical Association* **99**, 854–866.

Imbens, G. & Kalyanaraman, K. (2012), 'Optimal bandwidth choice for the regression discontinuity estimator', *The Review of Economic Studies* **79**, 933–959.

Imbens, G. W. (2000), 'The role of the propensity score in estimating dose-response functions', *Biometrika* **87**, 706–710.

Imbens, G. W. (2004), 'Nonparametric estimation of average treatment effects under exogeneity: A review', *The Review of Economics and Statistics* **86**, 4–29.

Imbens, G. W. (2010), 'Better late than nothing: Some comments on Deaton (2009) and Heckman and Urzua (2009)', *Journal of Economic Literature* **48**, 399–423.

Imbens, G. W. & Angrist, J. (1994), 'Identification and estimation of local average treatment effects', *Econometrica* **62**, 467–475.

Imbens, G. W. & Lemieux, T. (2008), 'Regression discontinuity designs: A guide to practice', *Journal of Econometrics* **142**, 615–635.

Imbens, G. W. & Wager, S. (2019), 'Optimized regression discontinuity designs', *Review of Economics and Statistics* **101**, 264–278.

Imbens, G. W. & Wooldridge, J. M. (2009), 'Recent developments in the econometrics of program evaluation', *Journal of Economic Literature* **47**, 5–86.

Kallus, N. (2017), 'Balanced policy evaluation and learning', *working paper, Cornell University* .

Keele, L. J. & Titiunik, R. (2015), 'Geographic boundaries as regression discontinuities', *Political Analysis* **23**, 127–155.

Kennedy, E. H., Ma, Z., McHugh, M. D. & 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.

Khan, S. & Tamer, E. (2010), 'Irregular identification, support conditions, and inverse weight estimation', *Econometrica* **78**, 2021–2042.

Kitagawa, T. (2015), 'A test for instrument validity', *Econometrica* **83**, 2043–2063.

Kitagawa, T. & Tetenov, A. (2018), 'Who should be treated? empirical welfare maximization methods for treatment choice', *Econometrica* **86**, 591–616.

Knaus, M., Lechner, M. & Strittmatter, A. (2018), 'Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence', *working paper, University of St. Gallen* .

Kolesár, M. & Rothe, C. (2018), 'Inference in a regression discontinuity design with a discrete running variable', *American Economic Review* **108**, 2277–2304.

Lalive, R. (2008), 'How do extended benefits affect unemployment duration? a regression discontinuity approach', *Journal of Econometrics* **142**, 785 – 806.

Landais, C. (2015), 'Assessing the welfare effects of unemployment benefits using the regression kink design', *American Economic Journal: Economic Policy* **7**, 243–278.

Lechner, M. (2001), Identification and estimation of causal effects of multiple treatments under the conditional independence assumption, *in* M. Lechner & F. Pfeiffer, eds, 'Econometric Evaluations of Active Labor Market Policies in Europe', Heidelberg: Physica.

Lechner, M. (2009), 'Sequential causal models for the evaluation of labor market programs', *Journal of Business and Economic Statistics* **27**, 71–83.

Lechner, M. (2010), 'The estimation of causal effects by difference-in-difference methods', *Foundations and Trends in Econometrics* **4**, 165–224.

Lechner, M., Miquel, R. & Wunsch, C. (2011), 'Long-run effects of public sector sponsored training in west germany', *Journal of the European Economic Association* **9**, 742–784.

Lechner, M. & Strittmatter, A. (2019), 'Practical procedures to deal with common support problems in matching estimation', *Econometric Reviews* **38**, 193–207.

Lee, D. (2008), 'Randomized experiments from non-random selection in u.s. house elections', *Journal of Econometrics* **142**, 675–697.

Lee, D. & Card, D. (2008), 'Regression discontinuity inference with specification error', *Journal of Econometrics* **142**, 655–674.

Lee, D. & Lemieux, T. (2010), 'Regression discontinuity designs in economics', *Journal of Economic Literature* **48**, 281–355.

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

Manski, C. F. (2004), 'Statistical treatment rules for heterogeneous populations', *Econometrica* **72**, 1221–1246.

McCrary, J. (2008), 'Manipulation of the running variable in the regression discontinuity design: A density test', *Journal of Econometrics* **142**, 698–714.

Mourifié, I. & Wan, Y. (2017), 'Testing late assumptions', *The Review of Economics and Statistics* **99**, 305–313.

Papay, J. P., Willett, J. B. & Murnane, R. J. (2011), 'Extending the regression-discontinuity approach to multiple assignment variables', *Journal of Econometrics* **161**, 203207.

Pearl, J. (2001), Direct and indirect effects, *in* 'Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence', Morgan Kaufman, San Francisco, pp. 411–420.

Porter, J. (2003), Estimation in the regression discontinuity model. mimeo.

Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N. H., Hastie, T. & Tibshirani, R. (2018), 'Some methods for heterogeneous treatment effect estimation in high dimensions', *Statistics in Medicine* **37**, 17671787.

Qian, M. & Murphy, S. A. (2011), 'Performance guarantees for individualized treatment rules', *Annals of Statistics* **39**, 11801210.

Robins, J. M. (1986), 'A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor effect', *Mathematical Modelling* **7**, 1393–1512.

Robins, J. M. & Greenland, S. (1992), 'Identifiability and exchangeability for direct and indirect effects', *Epidemiology* **3**, 143–155.

Robins, J. M., Hernan, M. A. & Brumback, B. (2000), 'Marginal structural models and causal inference in epidemiology', *Epidemiology* **11**, 550–560.

Robins, J. M., Mark, S. D. & Newey, W. K. (1992), 'Estimating exposure effects by modelling the expectation of exposure conditional on confounders', *Biometrics* **48**, 479–495.

Robins, J. M. & Rotnitzky, A. (1995), 'Semiparametric efficiency in multivariate regression models with missing data', *Journal of the American Statistical Association* **90**, 122–129.

Robins, J. M., Rotnitzky, A. & Zhao, L. (1994), 'Estimation of regression coefficients when some regressors are not always observed', *Journal of the American Statistical Association* **90**, 846–866.

Robins, J. M., Rotnitzky, A. & Zhao, L. (1995), 'Analysis of semiparametric regression models for repeated outcomes in the presence of missing data', *Journal of the American Statistical Association* **90**, 106–121.

Rosenbaum, P. R. & Rubin, D. B. (1983), 'The central role of the propensity score in observational studies for causal effects', *Biometrika* **70**, 41–55.

Rosenbaum, P. R. & Rubin, D. B. (1985), 'Constructing a control group using multivariate matched sampling methods that incorporate the propensity score.', *The American Statistician* **39**, 33–38.

Rothe, C. & Firpo, S. (2013), 'Semiparametric estimation and inference using doubly robust moment conditions', *IZA Discussion Paper No. 7564* .

Roy, A. (1951), 'Some thoughts on the distribution of earnings', *Oxford Economic Papers* **3**, 135–146.

Rubin, D. B. (1974), 'Estimating causal effects of treatments in randomized and nonrandomized studies', *Journal of Educational Psychology* **66**, 688–701.

Rubin, D. B. (1979), 'Using multivariate matched sampling and regression adjustment to control bias in observational studies', *Journal of the American Statistical Association* **74**, 318–328.

Rubin, D. B. (1990), 'Formal mode of statistical inference for causal effects', *Journal of Statistical Planning and Inference* **25**, 279–292.

Sant'Anna, P. H. C. & Zhao, J. B. (2018), 'Doubly robust difference-in-differences estimators', *working paper, Vanderbilt University* .

Sharma, A. (2016), 'Necessary and probably sufficient test for finding valid instrumental variables', *working paper, Microsoft Research, New York* .

Sianesi, B. (2004), 'An evaluation of the swedish system of active labor market programs in the 1990s', *The Review of Economics and Statistics* **86**, 133–155.

Simonsen, M., Skipper, L. & Skipper, N. (2016), 'Price sensitivity of demand for prescription drugs: Exploiting a regression kink design', *Journal of Applied Econometrics* **31**, 320–337.

Smith, J. & Todd, P. (2005), 'Rejoinder', *Journal of Econometrics* **125**, 365–375.

Stoye, J. (2009), 'Minimax regret treatment choice with finite samples', *Journal of Econometrics* **151**, 70–81.

Strezhnev, A. (2018), 'Semiparametric weighting estimators for multi-period difference-in-differences designs', *working paper, University of Pennsylvania* .

Tan, Z. (2006), 'Regression and weighting methods for causal inference using instrumental variables', *Journal of the American Statistical Association* **101**, 1607–1618.

Tchetgen Tchetgen, E. J. & Shpitser, I. (2012), 'Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis', *The Annals of Statistics* **40**, 1816–1845.

Thistlethwaite, D. & Campbell, D. (1960), 'Regression-discontinuity analysis: An alternative to the ex post facto experiment', *Journal of Educational Psychology* **51**, 309–317.

Tibshirani, R. (1996), 'Regresson shrinkage and selection via the lasso', *Journal of the Royal Statistical Society* **58**, 267–288.

Uysal, S. D. (2011), 'Doubly robust iv estimation of the local average treatment effects', *mimeo, University of Konstanz* .

van der Laan, M. & Rubin, D. (2006), 'Targeted maximum likelihood learning', *The International Journal of Biostatistics* **2**, 1–38.

Waernbaum, I. (2012), 'Model misspecification and robustness in causal inference: comparing matching with doubly robust estimation', *Statistics in Medicine* **31**, 1572–1581.

Wager, S. & Athey, S. (2018), 'Estimation and inference of heterogeneous treatment effects using random forests', *Journal of the American Statistical Association* **113**, 1228–1242.

Zhang, B., Tsiatis, A. A., Davidian, M., Zhang, M. & Laber, E. (2012), 'Estimating optimal treatment regimes from a classification perspective', *Stat* **1**, 103–114.

Zhao, Z. (2004), 'Using matching to estimate treatment effects: Data requirements, matching metrics, and monte carlo evidence', *Review of Economics and Statistics* **86**, 91–107.

Zhou, X., Mayer-Hamblett, N., Khan, U. & Kosorok, M. R. (2017), 'Residual weighted learning forestimating individualized treatment rules', *Journal of the American Statistical Association* **112**, 169–187.

Zhou, Z., Athey, S. & Wager, S. (2018), 'Offline multi-action policy learning: Generalization and optimization', *working paper, Stanford University* .

Zimmert, M. (2018), 'Efficient difference-in-differences estimation with high-dimensional common trend confounding', *working paper, University of St. Gallen* .

Zubizarreta, J. R. (2015), 'Stable weights that balance covariates for estimation with incomplete outcome data', *Journal of the American Statistical Association* **110**, 910–922.

Working Paper 504 august 2019

Author
Martin HUBER
University of Fribourg, Faculty of Economics and Social Sciences, Chair of Applied Econometrics - Evaluation of Public Policies, Bd. de Pérolles 90, 1700 Fribourg, Switzerland.
Phone: +41 26 300 8255; Email: martin.huber@unifr.ch; Website: http://www.unifr.ch/appecon/en/

Abstract
This chapter covers different approaches to policy evaluation for assessing the causal effect of a treatment or intervention on an outcome of interest. As an introduction to causal inference, the discussion starts with the experimental evaluation of a randomized treatment. It then reviews evaluation methods based on selection on observables (assuming a quasi-random treatment given observed covariates), instrumental variables (inducing a quasi-random shift in the treatment), difference-in-differences and changes-in-changes (exploiting changes in outcomes over time), as well as regression discontinuities and kinks (using changes in the treatment assignment at some threshold of a running variable). The chapter discusses methods particularly suited for data with many observations for a flexible (i.e. semi- or nonparametric) modeling of treatment effects, and/or many (i.e. high dimensional) observed covariates by applying machine learning to select and control for covariates in a data-driven way. This is not only useful for tackling confounding by controlling for instance for factors jointly affecting the treatment and the outcome, but also for learning effect heterogeneities across subgroups defined upon observable covariates and optimally targeting those groups for which the treatment is most effective.

Citation proposal
Martin Huber. 2019. «An introduction to flexible methods for policy evaluation». Working Papers SES 504, Faculty of Economics and Social Sciences, University of Fribourg (Switzerland)

Jel Classification
C21, C26, C29

Keywords
Policy evaluation, treatment effects, machine learning, experiment, selection on observables, instrument, difference-in-differences, changes-in-changes, regression discontinuity design, regression kink design.

Working Papers SES collection

Last published
497 Huber M., Solovyeva A.: On the sensitivity of wage gap decompositions; 2018
498 Isakov D., Pérignon C., Weisskopf J.-P.: What if dividends were tax-exempt? Evidence from a natural experiment; 2018
499 Denisova-Schmidt E., Huber M., Prytula Y.: The effects of anti-corruption videos on attitudes towards corruption in a Ukrainian online survey; 2019
500 Huber M.: A review of causal mediation analysis for assessing direct and indirect treatment effects. 2019
501 Arifine G., Felix R., Furrer O.: Multi-Brand Loyalty in Consumer Markets: A Qualitatively-Driven Mixed Methods Approach; 2019
502 Andre P., Delesalle E., Dumas C.: Returns to farm child labor in Tanzania; 2019
503 De La Rupelle M., Dumas C.: Health consequences of sterilizations; 2019

Catalogue and download links
http://www.unifr.ch/ses/wp
http://doc.rero.ch/collection/WORKING_PAPERS_SES

Publisher
Université de Fribourg, Suisse, Faculté des sciences économiques et sociales
Universität Freiburg, Schweiz, Wirtschafts- und sozialwissenschaftliche Fakultät
University of Fribourg, Switzerland, Faculty of Economics and Social Sciences

Bd de Pérolles 90, CH-1700 Fribourg
Tél.: +41 (0) 26 300 82 00
decanat-ses@unifr.ch www.unifr.ch/ses