Bodory, Hugo; Huber, Martin; Lechner, Michael

The Finite Sample Performance of Instrumental Variable-Based Estimators of the Local Average Treatment Effect When Controlling for Covariates

Computational Economics

Provided in Cooperation with:
Springer Nature

Suggested Citation: Bodory, Hugo; Huber, Martin; Lechner, Michael (2023) : The Finite Sample Performance of Instrumental Variable-Based Estimators of the Local Average Treatment Effect When Controlling for Covariates, Computational Economics, ISSN 1572-9974, Springer US, New York, NY, Vol. 64, Iss. 4, pp. 2053-2078, https://doi.org/10.1007/s10614-023-10507-y

Computational Economics (2024) 64:2053–2078
https://doi.org/10.1007/s10614-023-10507-y

The Finite Sample Performance of Instrumental Variable-Based Estimators of the Local Average Treatment Effect When Controlling for Covariates

Hugo Bodory[1] · **Martin Huber**[2] · **Michael Lechner**[3,4,5,6,7]

Accepted: 16 October 2023 / Published online: 14 November 2023

Abstract
This paper investigates the finite sample performance of a range of parametric, semi-parametric, and non-parametric instrumental variable estimators when controlling for a fixed set of covariates to evaluate the local average treatment effect. Our simulation designs are based on empirical labor market data from the US and vary in several dimensions, including effect heterogeneity, instrument selectivity, instrument strength, outcome distribution, and sample size. Among the estimators and simulations considered, non-parametric estimation based on the random forest (a machine learner controlling for covariates in a data-driven way) performs competitive in terms of the average coverage rates of the (bootstrap-based) 95% confidence intervals, while also being relatively precise. Non-parametric kernel regression as well as certain versions of semi-parametric radius matching on the propensity score, pair matching on the covariates, and inverse probability weighting also have a decent coverage, but are less precise than the random forest-based method. In terms of the average root mean squared error of LATE estimation, kernel regression performs best, closely followed by the random forest method, which has the lowest average absolute bias.

Keywords Instrumental variables · Local average treatment effects · Empirical Monte Carlo study

JEL Classification C21 · C26

1 Introduction

The evaluation of the causal effect of a treatment (e.g., fertility) on an outcome (e.g., labor supply) is frequently complicated by endogeneity, implying that the treatment is associated with unobserved characteristics affecting the outcome (e.g. personality traits, preferences, and values concerning family and working life). One may

Extended author information available on the last page of the article

nevertheless assess treatment effects in the presence of an instrumental variable (IV) which affects the treatment of (at least) some subjects in a monotonic way, does not directly affect the outcome (other than through treatment) and is as good as randomly assigned. Under these conditions, the local average treatment effect (LATE) on the compliers, the subpopulation whose treatment state reacts positively to the instrument, is identified, as discussed in Imbens and Angrist (1994), Angrist et al. (1996). In many empirical contexts, it may seem unlikely that the IV assumptions hold unconditionally, in particular when the treatment evaluation relies on observational data in which the instrument is not explicitly randomized like in an experiment. Depending on the application, it might, however, appear plausible that the IV assumptions hold conditional on covariates observed in the data. In this case, the LATE is identified and can be consistently estimated under certain conditions, see the discussions in Abadie (2003), Tan (2006), and Frölich (2007).

This paper assesses the finite sample performance of various parametric, semi-parametric, and non-parametric IV estimators when controlling for a fixed (i.e., pre-defined and low-dimensional) set of covariates by Monte Carlo simulations that are based on empirical labor market data from Angrist and Evans (1998). The latter study assesses the effect of fertility, defined as having at least three vs. two children, on mother's labor supply (for instance, a binary employment status or weeks employed per year), using twins at the second birth as instrument. The intuition for this IV strategy is that if a mother with one child get twins at the second birth, then fertility immediately increases to three rather than two children, implying a first stage effect of the twins instrument on the treatment. In the spirit of Huber et al. (2013), our empirical Monte Carlo simulation makes to a certain extent use of the empirical associations in the labour market data when assessing the various IV estimators, with the aim that our analysis is more closely linked to real world applications. We vary the simulation designs with respect to several dimensions, including treatment effect heterogeneity, instrument selectivity across observed covariates (namely age, race, and quarter of birth), instrument strength, the outcome distribution, and sample size. We analyse the performance of a range of estimators commonly considered in treatment and policy evaluation based on instruments, including two stage least squares, inverse probability weighting (IPW), matching, doubly robust estimation, and parametric as well as non-parametric regression based on the so-called Wald formula for LATE estimation.

We find that overall, non-parametric estimation based on the random forest, a machine learning algorithm controlling for covariates in a data-driven way, performs best in terms of coverage rates, which are defined as the share of simulations in which the true LATE is included in the 95% confidence interval of a LATE estimator. We note that the estimators' standard errors required for constructing confidence intervals are obtained by the non-parametric bootstrap, which naturally accounts for heteroscedasticity as well as uncertainty e.g. related to the first-step estimation of the instrument propensity scores and has performed very well in a simulation study by Bodory et al. (2020) on the variance estimation of treatment effect estimators. Furthermore, the random forest-based estimator is relatively precise, implying that the confidence interval is comparably short, which (conditional on having a decent coverage) appears desirable from the perspective of statistical power. The highest

overall precision has parametric regression based on the Wald formula, which has an acceptable coverage rate, too (albeit somewhat worse than the random forest-based estimator). Non-parametric kernel regression as well as certain versions of semi-parametric radius matching on the propensity score, pair matching on the covariates, and inverse probability weighting also have a decent coverage, but are less precise than the random forest-based method. Concerning the average root mean squared error of LATE estimation, kernel regression performs best (and also has the smallest average standard deviations), closely followed by the random forest method, which has the lowest average absolute bias. Overall, the random forest approach appears to be the (or among the) most favorable method(s) in terms of a combined assessment of coverage, precision, and model flexibility.

Our study contributes to a growing literature of simulation studies investigating the finite sample behavior of treatment effect estimators (such as IPW, matching, or doubly robust methods), see for instance (Frölich, 2004; Zhao, 2004; Lunceford & Davidian, 2004; Busso et al., 2014; Huber et al., 2013; Frölich et al., 2017). However, these previous studies focus on the selection-on-observables framework, implying that the treatment is exogenous (i.e., as good as random) conditional on covariates. The main contribution of the current study is that it appears to be the first empirical Monte Carlo simulation that offers a comprehensive analysis of the finite sample performance of a range of instrument-based estimators of the LATE, under the assumption that the instrument (rather than the treatment) is conditionally exogenous.

The remainder of this paper is organized as follows. Section 2 discusses the identifying assumptions for IV-based LATE evaluation in the presence of covariates. Section 3 introduces various parametric, semi-parametric, and non-parametric LATE estimators, as well as a bootstrap procedure for computing standard errors. Section 4 presents our empirical Monte Carlo simulation approach, namely the empirical data and the simulation designs. Section 5 presents the results on the finite sample performance of the LATE estimators. Section 6 concludes.

2 Identification of the LATE

In this section, we present the assumptions underlying the identification of the Local Average Treatment Effect (LATE) when controlling for covariates. To formalize the discussion, let us denote by D_i a possibly endogenous treatment received by unit i, and by Y_i the outcome variable based on which the treatment effect is to be evaluated. In their seminal paper, Imbens and Angrist (1994) define the LATE as the mean effect of Y_i in response to a change in D_i among the compliers, a subgroup whose D_i reacts to an exogenous shift in the instrumental variable, which is denoted by Z_i. To discuss the identification of the LATE, we make use of the potential outcomes framework introduced by Rubin (1974), which expresses causal effects as differences between potential outcomes under treatment and non-treatment. We adapt this concept to our instrumental variable setting with binary indicators D_i and Z_i, and define potential outcome and treatment variables for unit i in the following way:

$$Y_{i,z}^{d} = Y_i(D_i = d_i, Z_i = z_i) \tag{1}$$

$$D_{i,z} = D_i(Z_i = z_i), \tag{2}$$

with $d_i, z_i \in \{0, 1\}$. Using this framework, Angrist et al. (1996) show that units can be divided into two subgroups, compliers and noncompliers. Compliers are those induced to take the treatment when being assigned to it. Formally, this type of units is characterized by $D_{i,1} - D_{i,0} = 1$. The subgroup of noncompliers may consist of three further types, namely always-takers with $D_{i,1} = D_{i,0} = 1$, never-takers with $D_{i,1} = D_{i,0} = 0$, and defiers with $D_{i,1} - D_{i,0} = -1$. Note that the type of a single unit cannot be identified because the counterfactual potential treatment (that would have occurred under the alternative, rather than the factual instrument assignment) is not observed.

Abadie (2003), Tan (2006), and Frölich (2007) consider non-parametric LATE identification and estimation when controlling for observed covariates, denoted by X_i. We subsequently present the identifying assumptions in this context, which consist of (i) a monotonicity restriction on the treatment, (ii) the existence of compliers, (iii) conditional independence of the instrument and the share of compliance types, (iv) conditional mean independence of the outcome and the instrument, and (v) common support.

Assumption 1 (Monotonicity) $P(D_{i,0} > D_{i,1}) = 0$.

Assumption 2 (Existence of compliers) $P(D_{i,0} < D_{i,1}) > 0$.

Assumption 3 (Unconfounded type) $P(\tau_i = t|X_i = x_i, Z_i = 0) = P(\tau_i = t|X_i = x_i, Z_i = 1)$ for $t \in \{a, n, c\}$.

The types τ include always-takers a, never-takers n, and compliers c.

Assumption 4 (Conditional mean independence of the outcome) $E[Y_{i,Z_i}^0|X_i = x_i, Z_i = 0, \tau_i = t] = E[Y_{i,Z_i}^0|X_i = x_i, Z_i = 1, \tau_i = t]$ for $t \in \{n, c\}$,
$E[Y_{i,Z_i}^1|X_i = x_i, Z_i = 0, \tau_i = t] = E[Y_{i,Z_i}^1|X_i = x_i, Z_i = 1, \tau_i = t]$ for $t \in \{a, c\}$.

Assumption 5 (Common support) $Supp(X_i|Z_i = 1) = Supp(X_i|Z_i = 0)$.

Assumption 1 rules out the presence of defiers, a type whose treatment never complies with the instrument. Assumption 2 implies that the subgroup of compliers exists. Due to the conditional independence of the instrument and the shares of compliers, always-takers, and never-takers stated in Assumption 3, the first stage effect of the instrument on the treatment is identified conditional on covariates, such that any variables affecting both the instrument and the treatment are controlled for. The conditional mean independence in Assumption 4 rules out a direct average effect of the instrument on the outcome (exclusion restriction) and unobservables that jointly affect the instrument and the outcome when controlling for covariates. Finally,

Assumption 5 ensures that for all covariate values occurring in the population, either instrument value $Z_i \in \{0, 1\}$ exists such that the instrument is not deterministic in the covariates.

Under Assumptions 1 to 5, the LATE, denoted as $\theta = E[Y^1_{i,Z_i} - Y^0_{i,Z_i} | D_{i,1} - D_{i,0} = 1]$, is identified by

$$\theta = \frac{E_X[E[Y_i|Z_i = 1, X_i] - E[Y_i|Z_i = 0, X_i]]}{E_X[E[D_i|Z_i = 1, X_i] - E[D_i|Z_i = 0, X_i]]}. \tag{3}$$

Based on the insights of Rosenbaum and Rubin (1983), Frölich (2007) shows that identification is also obtained by conditioning on the instrument propensity score $p(x) := P(Z_i = 1|X_i = x)$ rather than the covariates, because it possesses the so-called 'balancing property'. That is, conditioning on the one-dimensional propensity score balances the distribution of the covariates across the states of the instrument. For this reason, the LATE is alternatively identified by

$$\theta = \frac{E_X[E[Y_i|Z_i = 1, p(X_i)] - E[Y_i|Z_i = 0, p(X_i)]]}{E_X[E[D_i|Z_i = 1, p(X_i)] - E[D_i|Z_i = 0, p(X_i)]]}. \tag{4}$$

3 Estimation and inference

In this section, we present parametric, semi-parametric, and non-parametric methods for estimating the LATE parameter θ introduced in Sect. 2. We also discuss a trimming rule that tackles limited common support in covariate values across instrument states, based on dropping observations which would obtain large weights in the estimator because their covariate values occur (almost) exclusively in only one of the instrument states. Finally, we provide an bootstrap procedure for estimating the standard errors of the LATE estimators.

3.1 Estimation

One method for the estimation of θ frequently applied in empirical work is two-stage least-squares (2SLS), which is easy to implement and computationally fast. However, the linearity assumption of the 2SLS estimator implies effect homogeneity, a restriction that may not hold in empirical studies. We consider 2SLS as a benchmark method, but also include more general LATE estimators that allow for effect heterogeneity of the LATE across values of the covariates.

Equations 3 and 4 imply that θ can be expressed as the ratio of two treatment effect estimators that account for covariate differences in the presence and absence of the instrument. The numerator gives the reduced form effect of Z_i on Y_i and the denominator the first stage effect of Z_i on D_i. Thus, a natural choice for the construction of estimators for θ is to substitute the expressions in the numerators and denominators of Eqs. 3 and 4 by estimators standardly applied in treatment or policy

evaluation, see for instance the surveys by Imbens (2004) and Imbens and Wooldridge (2009). Many treatment effect estimators are semi-parametric in the sense that (parametric) propensity score estimation is combined with non-parametric treatment effect estimation, using weighting, matching, or doubly robust methods. We consider such methods to estimate the LATE based on estimates of the instrument propensity score. We also vary the degree of flexibility of the estimators and implement parametric, semi-parametric, and non-parametric approaches to compute the reduced form and first stage effects in the numerators and denominators of Eqs. 3 and 4.

Smith and Todd (2005), among others, regard treatment effect estimators as weighted differences in outcomes. We apply this definition to the Wald formula and express the LATE as:

$$\hat{\theta} = \frac{\frac{1}{n_1}\sum_{i=1}^{n} z_i \hat{w}_i y_i - \frac{1}{n_0}\sum_{j=1}^{n}(1-z_j)\hat{w}_j y_j}{\frac{1}{n_1}\sum_{i=1}^{n} z_i \hat{w}_i d_i - \frac{1}{n_0}\sum_{j=1}^{n}(1-z_j)\hat{w}_j d_j}. \tag{5}$$

n denotes the size of an i.i.d. sample of realizations of $\{Y_i, D_i, Z_i, X_i\}$ with observation $i \in 1, ..., n$. $n_1 = \sum_{i=1}^{n} Z_i$ is the size of the subsample of those with $Z_i = 1$, $n_0 = n - n_1$, and $\hat{w}_i$ are weights that may depend on X_i or $\hat{p}(x)$, an estimate of the propensity score $p(x)$. Next, we discuss different methods of estimating $\hat{p}(x)$ and $\hat{w}_i$.

3.2 Instrument propensity scores

We consider two different approaches to balance the covariates across groups for units with $Z_i = 0$ and $Z_i = 1$. One is to directly control for covariates X_i, but some LATE estimators alternatively control for estimates of $p(x)$, which is motivated by the propensity score's balancing properties discussed in Rosenbaum and Rubin (1983). Their results imply that $p(x)$ is capable of equalizing the covariate distributions across instrument states, such that the instrument is conditionally independent of potential outcomes and treatments given the propensity score whenever independence holds conditional on the covariates. A practical advantage of controlling for the propensity score (rather than a vector of covariates) is that it is one-dimensional and thus, avoids the curse of dimensionality.

We compute $\hat{p}(x)$ in three different ways. Firstly, we specify a probit model to estimate the conditional probability $P(Z_i = 1|X_i = x_i)$ by

$$\hat{p}(x)^{probit} = \Phi(x_i^T \tilde{\beta}_{ML}), \tag{6}$$

where $\tilde{\beta}_{ML}$ denotes the estimated probit coefficients based on maximum likelihood and $\Phi(x_i^T \tilde{\beta}_{ML})$ is the cumulative distribution function of the standard normal distribution evaluated at $X_i^T \tilde{\beta}_{ML}$.

Secondly, we apply the covariate balancing propensity score (CBPS) method by Imai and Ratkovic (2014) to compute $\hat{p}(x)$. This methodology maximizes covariate balancing when predicting treatment assignment using the generalized method-of-moments (GMM) framework. Imai and Ratkovic (2014) show that the CBPS

method is robust to mild misspecifications of the propensity score model, which is estimated by the following expression:

$$\hat{p}(x)^{CBPS} = \Lambda(x_i^T \tilde{\beta}_{GMM}), \tag{7}$$

where $\tilde{\beta}_{GMM}$ are coefficients estimated by GMM and $\Lambda(x_i^T \tilde{\beta}_{GMM})$ is the cumulative distribution function of the standard logistic distribution evaluated at $x_i^T \tilde{\beta}_{GMM}$. We use the overidentified version of CBPS, with more moment conditions (based on the covariate balancing condition and the score of a logit model) than coefficients β_{GMM}, which are estimated by continuously updated GMM estimation:

$$\tilde{\beta}_{GMM} = \arg\min_{\beta} \bar{g}_{\beta}(Z, X)^T \Sigma_{\beta}(Z, X)^{-1} \bar{g}_{\beta}(Z, X). \tag{8}$$

$\bar{g}_{\beta}(Z, X)$ is the sample mean of the moment conditions and $\Sigma_{\beta}(Z, X)$ is a consistent variance estimator, described in more detail in Chapter 2.2 of Imai and Ratkovic (2014). See Heiler (2022) for a more detailed discussionof LATE estimation based on the CBPS.

Our third estimator of the instrument propensity score is fully non-parametric and based on kernel regression:

$$\hat{p}(x)^{lc} = \frac{\sum_{i=1}^{n} z_i K\left(\frac{x_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)}. \tag{9}$$

Equation 9 corresponds to the Nadaraya-Watson (local constant) kernel estimator, where K denotes the Epanechnikov kernel and bandwidth h is chosen by least-squares cross-validation, i.e., by minimizing the least squares cross validation error w.r.t. h, see Li and Racine (2006). As an alternative to using $\hat{p}(x)^{lc}$ as weighting function, we also apply the Nadaraya-Watson estimator for estimating the outcome and treatment models in Eq. 3, see our discussion on non-parametric estimation methods in Chapter 3.7.

A practically relevant issue of treatment effect methods is thin or lacking common support (or overlap) in the propensity score, which may compromise estimation due to a non-comparability across groups, see the discussions in Imbens (2004), Imbens and Wooldridge (2009), and Lechner and Strittmatter (2019). If specific propensity score values among one group are either very rare (thin common support) or absent (lack of common support) among the opposite group, as it may occur close to the boundaries of the propensity score, some units may receive a very large weight $\hat{w}_i$ in LATE estimation as provided in Eq. 5. In the case of thin common support, these observations could dominate the estimator of the LATE which may potentially entail an explosion of the variance. In the case of lacking common support, this even introduces asymptotic bias by giving a large weight to observations that are not comparable to observations in the opposite group in terms of the propensity score.

Huber et al. (2013) and Bodory et al. (2020) consider a trimming procedure to tackle common support issues in the sample also discussed in Imbens (2004), which

is asymptotically unbiased if common support holds asymptotically. It is based on setting the weights of those observations to zero whose relative share of all weights within either instrument state in Eq. 5 exceeds a particular threshold value in % (denoted by t):

$$\hat{w}_i = \hat{w}_i \mathbb{I}\left\{ \frac{z_i/\hat{p}(x)^{lc}}{\sum_{j=1}^{n} z_j/\hat{p}(x)^{lc}} + \frac{(1-z_i)/(1-\hat{p}(x)^{lc})}{\sum_{j=1}^{n}(1-z_j)/(1-\hat{p}(x)^{lc})} \leq t\% \right\}. \tag{10}$$

We set the threshold t to 5% and trim observations based on the weights of normalized IPW, see (3.3), irrespective of the LATE estimator considered. This changes (in finite samples) the target parameter due to discarding observations with extreme weights, but ensures common support prior to estimation. Note that our bootstrap variance estimators discussed in Sect. 3.8 account for the stochastic nature of trimming.

3.3 Inverse probability weighting (IPW)

Inverse probability weighting (IPW) reweighs (instrument) group-specific outcomes such that the distribution of the covariates in the total population is matched, see Hirano et al. (2003) for a more detailed discussion. We consider a normalized IPW estimator in our simulations, which performed well in several simulation studies on conditionally exogenous treatments, see for instance (Huber et al., 2013) and Busso et al. (2014). The IPW-based LATE estimator corresponds to

$$\hat{\theta}_{\text{IPW}} = \frac{\sum_{i=1}^{n} z_i y_i \left\{ \frac{\frac{1}{\hat{p}(x_i)}}{\sum_{j=1}^{n} \frac{z_j}{\hat{p}(x_j)}} \right\} - \sum_{i=1}^{n}(1-z_i) y_i \left\{ \frac{\frac{1}{1-\hat{p}(x_i)}}{\sum_{j=1}^{n} \frac{1-z_j}{1-\hat{p}(x_j)}} \right\}}{\sum_{i=1}^{n} z_i d_i \left\{ \frac{\frac{1}{\hat{p}(x_i)}}{\sum_{j=1}^{n} \frac{z_j}{\hat{p}(x_j)}} \right\} - \sum_{i=1}^{n}(1-z_i) d_i \left\{ \frac{\frac{1}{1-\hat{p}(x_i)}}{\sum_{j=1}^{n} \frac{1-z_j}{1-\hat{p}(x_j)}} \right\}}. \tag{11}$$

The normalizations $\sum_{j=1}^{n} \frac{z_j}{\hat{p}(x_j)}$ and $\sum_{j=1}^{n} \frac{1-z_j}{1-\hat{p}(x_j)}$ ensure that the weights in curly brackets add up to one. It is easy to see that (11) corresponds to (5) when setting $\hat{w}_i$ in the latter to $z_i n_1 \left\{ \frac{\frac{1}{\hat{p}(x_i)}}{\sum_{j=1}^{n} \frac{z_j}{\hat{p}(x_j)}} \right\} + (1-z_i) n_0 \left\{ \frac{\frac{1}{1-\hat{p}(x_i)}}{\sum_{j=1}^{n} \frac{1-z_j}{1-\hat{p}(x_j)}} \right\}$. IPW possesses the desirable property that it can attain the semiparametric efficiency bound (implying the smallest possible asymptotic variance) derived by Hahn (1998), if the propensity score is estimated non-parametrically (while this is generally not the case for parametric propensity scores). Furthermore, it is computationally inexpensive and easy to implement. However, evidence in the treatment effect literature suggests that IPW also has an important drawback: at the boundaries of the support of the propensity score, estimation may be unstable and the variance may explode in finite samples, see Frölich (2004) and Khan and Tamer (2010).

3.4 Doubly robust estimation

Doubly robust (DR) estimation combines IPW with outcome regression. It reweighs outcome models for different instrument states by the inverse of the propensity scores. Denoting the conditional mean outcomes in the presence and absence of the instrument by $\mu_z^y(x) := E[Y_i|Z = z_i, X_i = x_i]$ and $\mu_z^d(x) := E[D_i|Z_i = z_i, X_i = x_i]$, the DR LATE estimator corresponds to

$$\hat{\theta}_{\mathrm{DR}} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_1^y(x) + \frac{z_i(y_i-\hat{\mu}_1^y(x))}{\hat{p}(x)} - \hat{\mu}_0^y(x) - \frac{(1-z_i)(y_i-\hat{\mu}_0^y(x))}{1-\hat{p}(x)}\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}_1^d(x) + \frac{z_i(d_i-\hat{\mu}_1^d(x))}{\hat{p}(x)} - \hat{\mu}_0^d(x) - \frac{(1-z_i)(d_i-\hat{\mu}_0^d(x))}{1-\hat{p}(x)}\right)}. \tag{12}$$

For non-binary outcomes, we run OLS regression to compute $\hat{\mu}_z^y(x) = x_i^T\hat{\beta}_{z,OLS}$. For binary outcome and treatment variables, we apply probit regression to compute $\hat{\mu}_z(x) = \Phi(x_i^T\hat{\beta}_{z,ML})$. The coefficients β_z are estimated in the subgroups with $Z_i \in \{0, 1\}$. Differently to IPW, which exclusively relies on reweighing by the propensity score, the DR estimator remains consistent even if either $\hat{p}(x)$ or $\hat{\mu}_z(x)$ is misspecified, as it makes use of both, the treatment and outcome models. If both are correctly specified, the DR estimator is semi-parametrically efficient, as discussed in Robins et al. (1994).

3.5 Matching

Matching is based on assigning (matching) to each observation in one instrument state one or more units in the other instrument state with comparable covariates, in order to estimate the LATE based on the ratio of average differences in the outcome and the treatment across units with and without instrument in the matched sample. We implement multiple variants of two types of matching methods, pair and radius matching, to estimate θ.

Pair (or one-to-one) matching with replacement (implying that an observation may be matched several times) as discussed in Rubin (1973) matches to each reference observation exactly the observation with the most similar covariates in the opposite instrument state. This implies the following weights in Eq. 5:

$$\varpi_{i,j} = \mathbb{I}\left\{ |\hat{f}(x_i, x_j)| = \min_{k:Z_k\in\{0,1\}} |\hat{f}(x_i, x_k)| \right\}. \tag{13}$$

$\varpi_{i,j}$ is the weight of the outcome (or treatment) of observation j in one instrument group (e.g., $Z_j = 0$) when matched to unit i in the opposite group (e.g., $Z_i = 1$), with $Z_k = 1 - Z_i$. $\mathbb{I}\{\cdot\}$ is the indicator function, which is one if its argument is true and zero otherwise. $\hat{f}(\cdot)$ is a function of the difference in covariates between observations i and j. For example, the function could be defined as the difference in propensity score estimates of observations i and j in the case of propensity score matching or as a distance metric w.r.t. the covariate values of i and j like the Euclidean distance in the case of matching directly on the covariates. In pair matching, all weights are zero except for the observation j with the smallest difference with reference

unit i, which receives a weight of one. For propensity score matching, we base the weights on the distance of the one-dimensional propensity score, while for direct matching, we use a normalized Euclidean distance metric, where differences in the covariates are weighed by the inverse of the variances of X_i. Because only one observation is matched to each unit irrespective of the sample size and the potential availability of several suitable matches with similar covariates, pair matching is not efficient (i.e., does not attain the smallest possible variance asymptotically). On the other hand, it is likely more robust to propensity score misspecification than IPW, in particular if the misspecified propensity score model is only a monotone transformation of the true model, see for instance Zhao (2008), Millimet and Tchernis (2009), Waernbaum (2012), and Huber et al. (2013).

Radius matching as discussed in Rosenbaum and Rubin (1985) and Dehejia and Wahba (1999) uses *all* matches with propensity scores within a predefined radius around the reference unit, which trades off some bias in order to increase efficiency (or precision). This approach expectedly works relatively well if several comparable potential matches are available for a reference unit. In the simulations, we consider the radius matching algorithm of Lechner et al. (2011), which performed well in Huber et al. (2013), who also provide details on the radius matching-related weighting function $\hat{w}_i$ in Eq. 5. The estimator combines distance-weighted radius matching, where units within the radius are weighted proportionally to the inverse of their distance to the reference unit, with a regression-based bias correction, see Rubin (1979) and Abadie and Imbens (2011). For the bias correction, we apply an OLS regression adjustment for Y and a probit regression adjustment for D to remove small and large sample bias due to mismatches. Horowitz et al. (2014) provide a detailed description of the estimator. As in Lechner et al. (2011), the radius size in our simulations is defined as a function of the distribution of distances between reference units and matches in pair matching. Namely, it is set to 3 times the maximum pair matching distance. Note that we include radius matching both with and without conditioning on the covariate 'age at first birth' in addition to the propensity score to account for this influential confounder.

3.6 Parametric regression estimators

In our simulations, IPW, DR estimation, and matching are implemented with various degrees of flexibility in terms of parametric assumptions. We consider both semi-parametric versions based on parametric propensity score models, $\hat{p}(x)^{probit}$ and $\hat{p}(x)^{CBPS}$, as well as fully non-parametric estimators using the non-parametric propensity scores $\hat{p}(x)^{lc}$ (based on a local constant kernel regression) or when directly conditioning on X_i. For non-parametric DR estimation, also the conditional means of the binary treatment and binary (or non-binary) outcome $\hat{\mu}_z(x)$ are estimated by local constant (or local linear) kernel regressions.

In addition, we also consider several parametric treatment effect estimators. The first parametric approach computes the LATE by differences in the

conditional mean functions $\hat{\mu}_z(x)$, which are estimated by OLS regressions for non-binary outcomes and by probit regressions for the treatment and binary outcome variables (see Sect. 3.4). Formally, this regression-based LATE estimator corresponds to the following expression:

$$\hat{\theta}_{\text{REGR}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_1^y(x) - \hat{\mu}_0^y(x))}{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_1^d(x) - \hat{\mu}_0^d(x))}. \tag{14}$$

Furthermore, we apply two-stage least-squares (2SLS) estimation, which was also applied by Angrist and Evans (1998) for analysing the data our simulations are based on. 2SLS may be regarded as a benchmark method for instrumental variable estimation under the assumption of homogeneous treatment effects. Formally, the 2SLS estimator is given by

$$\begin{aligned}\hat{\theta}_{\text{2SLS}} = &\left[\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{x}_i^T\tilde{z}_i\right)\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{z}_i^T\tilde{z}_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{z}_i^T\tilde{x}_i\right)\right]^{-1} \\ &\times\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{x}_i^T\tilde{z}_i\right)\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{z}_i^T\tilde{z}_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{z}_i^T y_i\right),\end{aligned} \tag{15}$$

where $\tilde{x}_i := (1, x_{i,1}, \cdots, x_{i,K})$, $\tilde{z}_i := (\tilde{x}_i, z_i)$, and K denotes the number of covariates X_i. Note that in our just-identified settings with one treatment and one instrumental variable, the 2SLS estimator is numerically identical to the limited information maximum likelihood (LIML) estimator.

3.7 Further non-parametric estimators

We analyze the performance of three further non-parametric estimation methods that do not impose any functional form assumptions on the regression functions of the outcome or the treatment.

Firstly, we apply the generalized random forest (GRF) method, a non-parametric estimator introduced by Athey et al. (2019). GRF is a variant of random forest algorithms, a machine learning approach, see for instance the discussion in Lee et al. (2020) and citations therein. As described in Breiman (2001), random forests consist of averaging the predictions of many decision trees applied to different subsamples that are repeatedly drawn from the original data. In each of these samples, a decision tree partitions the space of X_i into a set of rectangles and computes the fitted value of Y_i as the average outcome in each of the rectangles. The partitions are chosen in a data-driven way such that the predictive performance is maximized (e.g. by minimizing the squared residuals based on the fitted values in each rectangle). A popular estimation algorithm for decision trees is CART (classification and regression tree), see for instance Chapter 9.2 in the textbook of Hastie et al. (2001).

GRF shares the core features of 'traditional' random forest algorithms like recursive partitioning, subsampling from the original data, and the random

selection of a subset of covariates at each partitioning step. However, as a methodological twist, GRF uses a gradient-based partitioning scheme and a particular (so-called 'honest') sample splitting technique (within any of the drawn sub-samples) that avoids overfitting the predictive models to the specificities of the data, see Wager and Athey (2018). Using the conditional expectation function $\mu_z(X_i)$ in Sect. 3.4 and applying the GRF to estimate the latter for the outcome and the treatment to obtain $\hat{\mu}^Y_{z,RF}(x)$ and $\hat{\mu}^D_{z,RF}(x)$ for $z \in \{0,1\}$ (where the subscript RF indicates the random forest approach), we compute the LATE as follows:

$$\hat{\theta}_{\text{RF}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}^Y_{1,RF}(x) - \hat{\mu}^Y_{0,RF}(x))}{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}^D_{1,RF}(x) - \hat{\mu}^D_{0,RF}(x))}. \tag{16}$$

Algorithm 1 in Athey et al. (2019) provides more details on the GRF method. We estimate the conditional expectations in Eq. 16 using the default options of the *causal_forest* function of the *grf* package for the statistical software *R*, see (Tibshirani et al., 2020). Alternatively, we could have estimated the predictions $\hat{\mu}_{z,RF}(x)$ by standard Breiman-type random forests (Breiman, 2001), which do, however, not safeguard against overfitting. In any case, it is important to note that our random forest approach making use of the Wald equation (16) is different to the IV-based causal forest suggested in Athey et al. (2019). The latter approach consists of first running random forests to obtain predictive models for the outcome, treatment, and instrument, respectively, as functions of covariates X and then residualizing the outcome, treatment, and instrument, i.e., purging their associations with the covariates based on the predictive models. Finally, the residualized variables are used in an IV regression to estimate the LATE. This approach is particularly attractive in high-dimensional contexts with many potential control variables in X, while our simulation study considers a low-dimensional setting with a fixed number of covariates.

In addition to the random forest, we consider non-parametric kernel regression for estimating the conditional mean functions $\hat{\mu}_z(x)$ defined in Sect. 3.4, see the subscript NP in the respective estimates in Eq. 17. For non-binary outcomes, $\hat{\mu}^y_{z,NP}(x)$ is estimated by local linear kernel regression, for the binary outcome and treatment variables, $\hat{\mu}^y_{z,NP}(x)$ and $\hat{\mu}^d_{z,NP}(x)$ are estimated by local constant kernel regression.

$$\hat{\theta}_{\text{NP}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}^y_{1,NP}(x) - \hat{\mu}^y_{0,NP}(x))}{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}^d_{1,NP}(x) - \hat{\mu}^d_{0,NP}(x))}. \tag{17}$$

Finally, we consider a (naive) LATE estimator that is based on the mean differences of the outcome and treatment variables, respectively, across instrument states, which in contrast to the other methods does not control for the covariates. Therefore, the consistency of this approach provided in Eq. 18 generally requires that the IV assumptions hold unconditionally, i.e., without conditioning on X.

Table 1 Point estimators

Estimators	Conditioning sets				
	Propensity scores			Covariates	None
	Probit	cbps	Local constant		
Inverse probability weighting	ipw^{probit}	ipw^{cbps}	ipw^{lc}		
Doubly robust	dr^{probit}	dr^{cbps}	dr^{lc}		
Pair matching	$pairmatch^{probit}$	$pairmatch^{cbps}$	$pairmatch^{lc}$	$pairmatch^{x}$	
Radius matching on propensity score	$radmatch^{probit}$	$radmatch^{cbps}$	$radmatch^{lc}$		
Radius matching on propensity score + covariate	$radmatchx^{probit}$	$radmatchx^{cbps}$	$radmatchx^{lc}$		
Parametric regressions				*reg*	
2SLS				*2sls*	
Random forests				*randforest*	
Non-parametric regressions				reg^{kernel}	
Mean differences (ignoring covariates)					*means*

$$\hat{\theta}_{\text{MEANS}} = \frac{\frac{1}{n_1}\sum_{i=1}^{n} y_i z_i - \frac{1}{n_0}\sum_{i=1}^{n} y_i(1-z_i)}{\frac{1}{n_1}\sum_{i=1}^{n} d_i z_i - \frac{1}{n_0}\sum_{i=1}^{n} d_i(1-z_i)}. \tag{18}$$

Table 1 summarizes the LATE estimators analysed in our simulation study along with the corresponding conditioning sets.

3.8 Inference

Treatment effect estimation frequently relies on the non-parametric bootstrap for statistical inference (Efron, 1979; Horowitz, 2001). In an extensive simulation study with a conditionally exogenous treatment, Bodory et al. (2020) find evidence that variance estimation of treatment effect estimators based on bootstrap procedures outperforms asymptotic variance approximations in terms of rejection and coverage probabilities in finite samples. These results even hold for matching estimators in small samples, despite the inconsistency of the non-parametric bootstrap for the (non-smooth) pair matching estimator, see the discussion in Abadie and Imbens (2008).

For this reason, we apply the non-parametric bootstrap to estimate the standard errors of all LATE estimators. This algorithm randomly draws B bootstrap samples of size n (the size of a simulation sample) with replacement out of each simulation sample and estimates the LATE in every draw. Denoting the B bootstrapped LATE estimators by $\hat{\theta}^b$, with $b \in \{1, 2, \ldots, B\}$, we estimate the standard error σ of a LATE estimator by

$$\hat{\sigma} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}^b - \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^b\right)^2}. \tag{19}$$

In line with Bodory et al. (2020), we set $B = 199$. Bootstrapping naturally accounts for heteroscedasticity as well as uncertainty due to trimming of influential observations and propensity score estimation.

4 Simulation design with empirical data

Simulations often rely on randomly generated data drawn from a probability distribution that is selected by the researcher. However, the data generating processes (DGPs) of such simulations may appear somewhat arbitrary in the sense that they might be far from reflecting typical associations between variables in empirical data. To improve upon this caveat, Huber et al. (2013) suggest a simulation design based on empirical data, also called Empirical Monte Carlo Study (EMCS), an idea that has been subsequently applied in several papers, see for instance Frölich et al. (2017), Huber et al. (2016), and Bodory et al. (2020), among others. Briefly, the idea of an EMCS is to randomly draw small samples from large real data sets while relying as much as possible on the empirical associations between the variables when generating the simulation designs.

Our study follows this EMCS approach to evaluate the properties of various IV estimators of the LATE, with the aim that the simulation designs are more closely linked to real world data. However, we point out that also in an EMCS, several important choices about the simulation features are to be made by the researcher such that the DGPs are not fully determined by the data, see the caveats raised by Advani and Sloczyński (2013). The remainder of this section describes the implementation of our EMCS. We first present the empirical labor market data underlying our simulations and then provide the steps for generating the various simulation designs.

4.1 Database

Our simulations are based on empirical data analysed in Angrist and Evans (1998), who aim at exploiting exogenous variation in family size to evaluate the treatment effect of fertility, defined as having at least three vs. two children, on female labor supply. This database is well suited to analyze the finite sample properties of IV estimators by means of an EMCS for several reasons. First, the data set is large, as it comprises 394840 observations and therefore easily allows one to draw many different random subsamples. Furthermore, the data contains a strong instrument that

importantly affects fertility, namely twins at second birth.[1] Finally, it provides demographic information on the mothers, which may be used as covariates to control for potential confounders of the instrument and the outcome.

Coming from the 1980 Census Public Use Micro Samples (PUMS), the data set contains information on young mothers aged 21 to 35, all of which gave birth to at least two children. Our analysis considers two different outcomes, the number of weeks worked within one year (with 43% zeros) and an indicator for being employed at all in that year. The binary treatment variable indicates if a mother has more than two kids (treatment is one) or two kids (treatment is zero). The binary instrumental variable is one if a mother gave birth to twins at second birth and zero otherwise. The covariates considered in our simulation include mother's age, mother's age at first birth, race, and quarter of birth.

Table 2 reports descriptive statistics of the database, by treatment indicator (more than two kids) and the instrument (twins at second birth). The upper part presents descriptives for the two labor market outcomes 'weeks worked' (in weeks) and 'worked for pay' (binary). There are large differences between the outcomes of the treated and non-treated in terms of the standardized difference statistic as suggested by Rosenbaum and Rubin (1985) (the literature considers values around 20 and above as severely unbalanced). The line underneath the outcomes in Table 2 gives details on the treatment variable. Not surprisingly, the treatment fully complies with the instrument if the latter equals one, because all mothers with twins at second birth ($Z_i = 1$) necessarily have more than two children ($D_i = 1$). The subsequent row of Table 2 provides information on the instrument. It reveals that 2% of women with at least three children have twins at their second birth. Considering the covariates, the standardized differences show that mothers' characteristics are partly unbalanced across treatment states, whereas they are well balanced across instrument states, in line with a randomly assigned instrument. The randomness of the instrument is also supported by the pseudo-R2 statistic with a value of 0.2% when regressing the instrument on the covariates.

4.2 Simulation designs

Data generating processes (DGPs) may differ in (infinitely) many dimensions. We select ten practically relevant dimensions for varying the specifications of our simulation models. These dimensions include: effect homogeneity vs. heterogeneity, randomness vs. non-randomness of the instrument, varying levels of instrument strength, binary vs. non-binary outcome distributions, and different sample sizes. Summary statistics of all DPGs are presented in Table 3.

We start by assuming homogeneous treatment effects with a randomly assigned instrument and the empirically observed instrument strength. To evaluate the performance of the estimators under these conditions, we define a new *population*

[1] There may be cases where the randomness of twin births is violated, see Farbmacher et al. (2018) for a discussion on dizygotic twinning. In our simulation study, we artificially generate random and non-random instrument assignments.

Table 2 Descriptive statistics of the full sample

Variables	More than two kids					Twins at second birth				
	No		Yes		St.diff	No		Yes		St.diff
	Mean	Std	Mean	Std		mean	Std	Mean	Std	
Outcomes										
Weeks worked / 10	2.32	2.26	1.73	2.13	19	2.09	2.23	1.89	2.19	6
Worked for pay (binary)	0.61		0.49		17	0.57		0.52		7
Treatment										
More than two kids (binary)						0.40		1		123
Instrument										
Twins at second birth (binary)	0		0.02		15					
Covariates: Mothers' characteristics										
Age / 10	2.98	0.36	3.05	0.34	14	3.01	0.35	3.04	0.34	5
Age at first birth /10	2.06	0.30	1.94	0.27	29	2.01	0.29	2.04	0.30	5
African American (binary)	0.10		0.15		11	0.12		0.15		6
Other race (binary)	0.18		0.18		1	0.18		0.20		3
First quarter of birth (binary)	0.24		0.24		1	0.24		0.24		1
Second quarter of birth (binary)	0.24		0.25		1	0.24		0.24		0
Third quarter of birth (binary)	0.27		0.27		0	0.27		0.27		0
Fourth quarter of birth (binary)	0.25	0.43	0.25	0.43	0	0.25	0.43	0.26	0.44	1
$X_i\tilde{\beta}$	−0.35	0.36	−0.14	0.36	42	−2.39	0.05	−2.38	0.05	10
$\Phi(X_i\tilde{\beta})$	0.37	0.13	0.45	0.14	42	0.01	0	0.01	0	10
Number of obs., Pseudo-R^2 in %	236089		158751		10.5	391460		3380		0.2

Notes: The statistics mean, std, and st.diff stand for mean, standard deviation, and standardized difference in percent, respectively. The standardized difference is defined as the absolute difference of means normalized by the square root of the sum of estimated variances of the particular variables in both subsamples (see e.g. 2009, p. 24). $\tilde{\beta}$ denotes the estimated probit coeffcients and $\Phi(X_i\tilde{\beta})$ is the cumulative distribution function of the standard normal distribution evaluated at $X_i\tilde{\beta}$. Pseudo-R^2 is the so-called Nagelkerke's R^2: $(1 - exp(-(-2(l_0 - l_1))/n)/(1 - exp(-(-2l_0)/n)$, where l_0 and l_1 are the log likelihoods for the null and full model, respectively, and n denotes the number of observations

Table 3 Summary statistics (DGPs)

Random selection Twins-2	Empirically observed strength Twins-2	Share (%) More than 2 children	Share (%) Twins-2	St.diff. (%) Twins-2 $P(Z\|X)$	Pseudo-R^2 (%) Twins-2 $P(Z\|X)$	First stage (%) Twins-2	LATE		Trimming (%)	
							1000	2000	1000	2000
Effect homogeneity (Weeks worked)										
Yes	Yes	70	50	23	4	60	0	0	0.03	0.02
Yes	No	43	50	23	4	6	0	0	0.03	0.02
No	Yes	70	50	51	16	61	0	0	0.06	0.03
No	No	43	50	51	16	7	0	0	0.06	0.03
Effect homogeneity (Worked for pay)										
Yes	Yes						0	0		
Yes	No						0	0		
No	Yes						0	0		
No	No						0	0		
Effect heterogeneity (Weeks worked)										
Yes	Yes	71	50	23	4	57	− 0.91	0.27	0.03	0.02
Yes	No	85	50	23	4	30	0.52	0.51	0.03	0.02
No	Yes	71	50	51	16	58	− 0.09	− 0.09	0.06	0.05
No	No	82	50	51	16	34	− 0.21	− 0.21	0.06	0.05
Effect heterogeneity (Worked for pay)										
Yes	Yes						0.00	0.00		
Yes	No						0.01	0.01		
No	Yes						− 0.01	− 0.01		
No	No						− 0.02	− 0.02		

Note: The numbers in columns 3-9 are averages over all simulation replications with samples of sizes 1000 and 2000. 'St.diff': Standardized difference defined as absolute mean difference normalized by the square root of the sum of the estimated variances of the particular variables in both subsamples [Imbens and Wooldridge (2009), p. 24]. 'std': standard deviation. Pseudo-R^2 is the so-called Nagelkerke's R^2: $(1 - exp(-(-2(l_0 - l_1))/n)/(1 - exp(-(-2l_0)/n)$, where l_0 and l_1 are the log likelihoods for the null and full model, respectively, and n is the sample size. Trimming: Share of dropped units due to the trimming rule in Eq. 10 or non-convergence of propensity scores. All statistics that do not dependend on the outcome variables are only presented for Weeks worked

for which the true LATE is equal to zero. To this end, we drop all 3380 observations from the database who receive the instrument ($Z_i = 1$). Among the remaining 391460 observations with instrument state $Z_i = 0$ (no twins at second birth), there is no reduced form effect of the instrument on the outcome or first stage effect of the instrument on the treatment, such that there exists no LATE. After that, we create a pseudo-instrument and artificially assign $Z_i = 1$ to those who are similar to the 3380 discarded observations in terms of observed characteristics. This similarity is determined by $1 : M$ matching on the covariates without replacement. By setting $M = 58$, we assign $Z_i = 1$ to approximately half of the observations, see column 4 of Table 3. In addition, we set the treatment state of everyone with $Z_i = 1$ to $D_i = 1$ (as in the original database) to maintain the empirically observed instrument strength. Finally, we draw small samples from our new *population* to compare the finite sample properties of alternative LATE estimators.

To simulate specifications with a weaker instrument, we reduce the first stage effect by lowering the impact of Z_i on D_i. Instead of setting all observations with $Z_i = 1$ to $D_i = 1$, we change the treatment status from zero to one only for those with $Z_i = 1$ for which the condition $D_i = \mathbb{1}(u_i > 1.25)$ holds. $\mathbb{1}(\cdot)$ denotes the indicator function which is one if its argument is true, otherwise it is zero, and u_i is a standard normally distributed random variable. Column 7 of Table 3 displays the first stage coefficients for the different DGPs.

The randomness of the instrument implies that the covariates are balanced across groups. To mimic a non-random assignment of the instrument, we increase the magnitude of instrument selectivity in the following way. We first estimate the propensity score $\hat{p}(1.5X_i)^{probit}$ (see Eq. 6) using the original database. Then, we change the instrument status Z_i from zero to one for observations with characteristics similar to the 3380 observations dropped from the original database (with $Z_i = 1$). We obtain such similar matches by $1 : M$ matching on the estimated propensity score $\hat{p}(1.5X_i)^{probit}$, with $M = 22$. Next, we assign $D_i = 1$ to all observations with $Z_i = 1$. Based on this modified data set, we increase the selection into the instrument by discarding the best matches for the newly created observations with $Z_i = 1$ among observation with $Z_i = 0$. To find the best matches to be discarded, we apply $1 : M$ matching on a newly estimated propensity score $\hat{p}(X_i)^{probit}$ (with the modified instrument assignments), where $M = 3$. The selectivity of the instrument is provided in columns 5 and 6 in Table 3.

To model a scenario with non-constant treatment effects, we introduce effect heterogeneity with respect to age and race as follows. We add to the existing control variables squared and cubic terms of both age variables ('age' and 'age at first birth') and interact the unmodified age variables with the indicator variable for African Americans. This new set of control variables for settings with effect heterogeneity is denoted by X_i^{het} for each unit i. We generate Y_i and D_i in each simulation sample according to the rules $Y_i = Y_{i,1}^d Z_i + Y_{i,0}^d (1 - Z_i)$ and $D_i = D_{i,1} Z_i + D_{i,1}(1 - Z_i)$. To this end, we compute the non-binary potential outcomes based on the equation $y_{i,z}^d = x_i^{het} \hat{\beta}_{OLS} + \hat{\sigma} v_i$, where v_i is a standard normally distributed random variable. $\hat{\beta}_{OLS}$ and $\hat{\sigma}$ are the coefficients and residual standard deviation of OLS regressions in subsamples by instrument state $Z_i \in \{0, 1\}$ of our new *population*. The binary potential

Table 4 Coverage rates and intervals

Point estimators	Coverage rates	diff (pp)	Confidence intervals	diff (%)
randforest	95.0	0.0	228.8	8.0
$radmatch^{probit}$	95.1	0.1	356.5	68.3
reg^{kernel}	94.8	0.2	329.4	55.5
$pairmatch^{x}$	94.6	0.4	265.6	25.4
ipw^{cbps}	95.5	0.5	290.4	37.1
$radmatchx^{cbps}$	94.4	0.6	284.5	34.3
$radmatchx^{probit}$	94.1	0.9	265.6	25.4
reg	96.0	1.0	211.8	0.0
2*sls*	96.0	1.0	260.9	23.2
dr^{probit}	96.0	1.0	257.5	21.6
means	90.5	4.5	234.5	10.7

Notes: 'diff' indicates the difference to the left best performer in percentage points (pp) or in percent (%)

outcomes are computed based on $Y^d_{i,z} = \mathbb{1}(X_i^{het}\hat{\beta}_{probit} + v_i > 0)$, where $\mathbb{1}(\cdot)$ is the indicator function and $\hat{\beta}_{probit}$ are the coefficients estimated from probit models in subsamples by instrument state of our new *population*. The potential treatments $D_{i,1}$ are set to one, whereas $D_{i,0}$ is computed analogously to $Y^d_{i,z}$ in the binary outcome case.

We combine these variations in the DGPs with respect to effect heterogeneity, instrument strength, and instrument selectivity with smaller and larger sample sizes of 1000 and 2000, respectively, and with binary and non-binary outcome distributions. We run 2000 simulations for the smaller and 1000 simulations for the larger samples. Table 3 presents summary statistics of the DGPs considered in our simulation study.

5 Results

This section presents results about the finite sample performance of various LATE estimators across different DGPs. We rank the estimators by their coverage rates, which are defined as the share of simulations in which the true LATE is included in the 95% confidence interval of the respective LATE estimator. We recall that the standard errors for computing those confidence intervals come from the non-parametric bootstrap, as discussed in Sect. 3.8. For the sake of brevity, we subsequently only discuss a selection of our results, which conveys the main message of our findings. In the Appendix, we include more detailed results.

Table 4 provides the average coverage rates and lengths of confidence intervals across all DGPs of any parametric, semi-parametric, or non-parametric LATE estimator which performs best (in terms of coverage) in at least one of the ten DGPs discussed in Sect. 4.2. We find that only the non-parametric random forest-based LATE estimator described in Eq. 16 attains exactly the nominal coverage size of

Table 5 Average absolute biases, standard deviations, and root mean square errors

Point estimators	Bias effect	diff (%)	sd	diff (%)	rmse	diff (%)	Bias se	diff (%)
randforest	0.6	0.0	8.0	14.1	8.0	11.9	1.7	0.0
$radmatch^{probit}$	1.3	100.1	12.0	70.0	12.0	67.3	10.8	540.7
reg^{kernel}	1.4	124.8	7.0	0.0	7.2	0.0	4.2	149.3
$pairmatch^{x}$	0.8	21.4	9.9	40.3	9.9	37.6	5.8	247.4
ipw^{cbps}	1.2	85.4	7.4	5.4	7.5	4.7	4.4	162.4
$radmatchx^{cbps}$	1.1	76.9	8.6	22.9	8.7	21.5	4.1	142.8
$radmatchx^{probit}$	1.2	86.4	9.7	38.1	9.8	36.4	3.4	100.1
reg	0.7	10.1	7.3	3.7	7.3	1.9	4.0	140.0
2*sls*	0.7	11.3	7.9	12.7	7.9	10.6	3.8	124.0
dr^{probit}	0.8	31.9	7.6	8.6	7.7	6.9	3.9	133.3
means	3.4	440.1	7.9	11.7	8.8	21.9	5.6	235.3

Notes: 'bias effect' denotes the absolute bias from the true treatment effect, 'sd' is the standard deviation of the estimator, 'rmse' stands for root mean squared error, and 'bias se' indicates the median bias of the estimated bootstrap standard error. 'diff' indicates the difference to the left best performer in percent (%)

95% on average. Furthermore, its average length of confidence intervals is the second shortest among the estimators analyzed in Table 4, 8% larger than the average interval of the parametric regression estimator, the nominal size of which is 96%. Conditional on obtaining a decent coverage, a short confidence interval is desirable in terms of precision, as it implies a lower estimation uncertainty. Three out of the four LATE estimators whose average coverage rates come closest to 95% are non-parametric, with those of non-parametric kernel regression (94.8%) and pair matching on the covariates (94.6%) having a minor under-coverage. Also semi-parametric radius matching on the propensity score performs decent in terms of coverage rates, with the probit-based version attaining an average rate of 95.1%, and two further versions achieving 94.4% and 94.1%, respectively. Furthermore, IPW using the CBPS method for propensity score estimation reaches a satisfactory average coverage rate, too, namely 95.5%.

In terms of average coverage, one might argue that all estimators in Table 4 but the mean differences estimator (which ignores covariates) perform decently, with coverage distortions amounting to at most one percentage point among methods controlling for covariates. In terms of the average length of the confidence intervals, however, parametric regression and the random forest clearly outperform radius matching, kernel regression, pair matching, IPW, 2SLS, and DR estimation, whose intervals are substantially longer (by 22–68%). According to the results in Table 4, the random forest-based LATE estimator is the preferred choice when paying relatively more attention to coverage (and sacrificing some precision compared to parametric regression), while parametric regression is the preferred choice for maximizing precision (when accepting average coverage distortions of 1 percentage point).

The coverage accuracy of the different estimation methods is related to the bias and variance of the LATE estimators, as well as the bias of the bootstrap-based

Table 6 Best performing estimators in terms of coverage rates

DGP feature	Point estimator			Coverage (best est.)
	Parametric	Semi-parametric	Non-parametric	
Effect homogeneity	2*sls*	**radmatchx$^{\text{probit}}$**	*means*	96.0
Effect heterogeneity	**2sls**	*dr^{probit}*	*randforest*	94.9
Standard selection	*reg*	*radmatchx^{cbps}*	**randforest**	95.0
Strong selection	2*sls*	*ipw^{cbps}*	**randforest**	95.1
Weaker first stage	*reg*	**dr$^{\text{probit}}$**	*randforest*	94.8
Strong first stage	*reg*	**radmatchx$^{\text{probit}}$**	*pairmatch^{x}*	95.9
Non-binary outcome	*reg*	**radmatch$^{\text{probit}}$**	*randforest*	95.0
Binary outcome	2*sls*	**radmatch$^{\text{probit}}$**	*randforest*	95.1
Small sample size	*reg*	*radmatch^{probit}*	**reg$^{\text{kernel}}$**	94.8
Larger sample size	2*sls*	**radmatch$^{\text{probit}}$**	*reg^{kernel}*	94.9

Notes: The in terms of average coverage best performing LATE estimator given a specific DGP feature is shown in bold print and its average coverage is reported in the last column

standard error. Table 5 provides details on these statistics. We find that two non-parametric methods perform best when considering the bias of LATE estimation, the standard deviation, and the mean squared error (i.e., the sum of the squared bias and the variance), as well as the bias of the standard error (relative to a LATE estimator's true standard deviation). The random forest-based LATE estimator has on average the smallest deviation from the true LATE, with its absolute bias amounting to 0.6. The non-parametric kernel regression estimator has the smallest average standard deviation among the estimators in Table 5, amounting to 7.0. It also performs best in terms of root mean squared errors across DGPs with an average value of 7.2. When considering the median bias of the bootstrap standard errors relative to the true standard deviations of the respective LATE estimators, the inference method of the random forest-based LATE estimator performs best, with an average median bias of 1.7. The averages of the median biases of its competitors are on average at least 100% larger.

Our findings suggest that the coverage accuracy is mainly driven by a LATE estimator's bias. This is for instance the reason why the mean differences estimator (which ignores covariates), whose bias exceeds that of the random forest-based LATE estimator by 440.1%, shows a relatively poor coverage in Table 4. Also the OLS estimator performs poorly in terms of coverage, due to its high bias, while its variance is small (results not presented but available on request). The performance of all LATE estimators by DGP is presented in Section A.3 of the Appendix. Tables 18–49 provide details on the coverage rates, biases, standard deviations, and root mean squared errors.

Table 6 lists the best performing LATE estimators in terms of average coverage, separately for each of the ten DGP features (see the rows) as well as for parametric, semi-parametric, and non-parametric methods (see the columns). The in terms of average coverage best performing LATE estimator given a specific DGP feature is

shown in bold print and its average coverage is reported in the last column of the table. The results suggest that semi-parametric and non-parametric estimators come closest to the nominal coverage rate of 95% and for any of these best performing estimators, the size distortion is at most one percentage point. The radius matching algorithm of Lechner et al. (2011) (with or without controlling for a covariate in addition to the propensity score) most frequently performs best both among the semi-parametric LATE estimators (in 70% of cases) and overall (in 50% of cases). Radius matching achieves the best average coverage in settings with effect homogeneity, a strong instrument, non-binary and binary outcomes, and under a larger sample size. For specifications with standard and strong selection into the instrument, the non-parametric random forest-based estimator is closest to the nominal size. Considering scenarios with effect heterogeneity, a weaker instrument, and a small sample size, the best performers are LATE estimators based on 2SLS, DR, and non-parametric regression, respectively.

The performance of the best performing LATE estimators across the ten DGP features is presented in Section A.2 of the Appendix. Tables 8–17 provide information on the average coverage rates, biases, standard deviations, and root mean squared errors. When taking precision into account, the random forest-based approach appears very competitive relative to other semi- or non-parametric methods like radius matching or IPW, as it tends to have substantially shorter confidence intervals while still attaining very decent coverage rates (even under DGP features where other estimators slightly dominate in terms of coverage).

6 Conclusion

This paper presented a simulation study based on empirical labor market data to investigate the finite sample properties of a range of point estimators of the local average treatment effect (LATE) when controlling for a fixed (and low-dimensional) set of covariates. The structure of these estimators is inspired by the Wald estimator, consisting of the ratio of the estimated reduced form effect of the instrument on the outcome and the estimated first stage effect of the instrument on the treatment. Furthermore, we applied the non-parametric bootstrap to estimate the standard errors and the 95% confidence intervals of the LATE estimators. We find that among the LATE estimators considered, non-parametric kernel regression has the smallest average root mean squared error across the different simulations, closely followed by the random forest-based approach, which has the lowest average absolute bias. The random forest method also performs very competitive in terms of average coverage rates, while at the same time having relatively narrow confidence intervals, which is attractive in terms of precision. Specific versions of semi-parametric radius matching on the propensity score, nonparametric kernel regression, inverse probability weighting, and pair matching on the covariates perform decently in terms of coverage, too, but have substantially wider confidence intervals. Parametric regression based on the Wald estimator is most precise, while also having decent coverage (albeit less so than the random forest method). Overall, the random forest-based

estimator seems to be the (or among the) most attractive method(s) in terms of a combined assessment of coverage, precision, and model flexibility.

While our Monte Carlo study is quite comprehensive in terms of simulation designs considered (with varying treatment effect heterogeneity, instrument selectivity and strength, outcome distributions, and sample sizes), we point out that our results do not necessarily generalize to simulation or empirical frameworks that are very different from the scenarios considered in this study. One interesting domain future simulation studies might want to consider are high-dimensional settings or 'big data' contexts in which many potential covariates are available as control variables. Such simulation designs would facilitate the assessment of LATE estimators that are particularly tailored to such high-dimensional scenarios, by combining instrument-based causal analysis with machine learning methods for selecting (important) control variables in a data-driven way. This concerns in particular so-called (Neyman, 1959)-orthogonal estimators, which are robust to moderate biases in the estimation of instrument propensity scores and conditional mean outcomes. Examples for such methods are double/debiased machine learning (Chernozhukov et al., 2018), causal forest algorithms as suggested by Athey et al. (2019) and Lechner and Mareckova (2022), or R-learning (Nie & Wager, 2020), which have not been considered in the low-dimensional context of this study.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10614-023-10507-y.

Author Contributions All authors contributed to the conception and design of this simulation study.

Funding Open access funding provided by University of St.Gallen. The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability The research idea was developed by Michael Lechner. Data processing and the implementation of the simulations were performed by Hugo Bodory and Martin Huber.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Ethical Approval The first draft of the manuscript was written by Hugo Bodory and all authors contributed to and approved the final version of the manuscript.

References

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

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 and Economic Statistics, 29*, 1–11.

Advani, A., & Sloczyński, T. (2013). Mostly harmless simulations? On the internal validity of empirical Monte Carlo studies. IZA Discussion Paper No. 7874.

Angrist, J., & Evans, W. (1998). Children and their parents labor supply: Evidence from exogeneous variation in family size. *American Economic Review, 88*, 450–477.

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

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics, 47*(2), 1148–1178.

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

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

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.

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*(1), C1–C68.

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.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*, 1–26.

Farbmacher, H., Guber, R., & Vikström, J. (2018). Increasing the credibility of the Twin birth instrument. *Journal of Applied Econometrics, 33*(3), 457–472.

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

Frölich, M. (2007). Nonparametric IV estimation of local average treatment effects with covariates. *Journal of Econometrics, 139*(1), 35–75.

Frölich, M., Huber, M., & Wiesenfarth, M. (2017). The finite sample performance of semi- and non-parametric estimators for treatment effects and policy evaluation. *Computational Statistics and Data Analysis, 115*, 91–102.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica, 66*(2), 315–331.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). *The elements of statistical learning*. Springer Series in Statistics. Springer New York Inc.,.

Heiler, P. (2022). Efficient covariate balancing for the local average treatment effect. *Journal of Business & Economic Statistics*, *40*(4), 1569–1582. https://doi.org/10.1080/07350015.2021.1946067

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

Horowitz, J. L. (2001). The Bootstrap. In J. J. Heckman & E. Learer (Eds.), *Handbook of Econometrics* (pp. 3159–3228). North-Holland.

Huber, M., Lechner, M., & Mellace, G. (2016). The finite sample performance of estimators for mediation analysis under sequential conditional independence. *Journal of Business & Economic Statistics, 34*(1), 139–160.

Huber, M., Lechner, M., Steinmayr, A. (2014). Radius matching on the propensity score with bias adjustment: Tuning parameters and finite sample behaviour. forthcoming in Empirical Economics.

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

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

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., & Angrist, J. (1994). Identification and estimation of local average treatment effects. *Econometrica, 62*, 467–475.
Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature, 47*, 5–86.
Khan, S., & Tamer, E. (2010). Irregular identification, support conditions, and inverse weight estimation. *Econometrica, 78*, 2021–2042.
Lechner, M., & Mareckova, J. (2022). "Modified Causal Forest," working paper, University of St. Gallen, School of Economics and Political Science.
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*(2), 193–207.
Lee, T.-H., Ullah, A., & Wang, R. (2020). *Bootstrap aggregating and random forest* (pp. 389–429). Springer.
Li, Q., & Racine, J. S. (2006). *Nonparametric econometrics: Theory and practice, no. 8355 in economics books*. Princeton University Press.
Lunceford, J. K., & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine, 23*, 2937–2960.
Millimet, D., & Tchernis, R. (2009). On the specification of propensity scores, with applications to the analysis of trade policies. *Journal of Business & Economic Statistics, 27*, 297–315.
Neyman, J. (1959). *Optimal asymptotic tests of composite statistical hypotheses* (pp. 416–444). Wiley.
Nie, X., & Wager, S. (2020). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika, 108*, 299–319.
Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association, 89*, 846–866.
Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 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.
Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics, 29*, 159–183.
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.
Smith, J., & Todd, P. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics, 125*, 305–353.
Tan, Z. (2006). Regression and weighting methods for causal inference using instrumental variables. *Journal of the American Statistical Association, 101*, 1607–1618.
Tibshirani, J., Athey, S., Friedberg, R., Hadad, V., Hirshberg, D., Miner, L., Sverdrup, E., Wager, S., & Wright, M. (2020). GRF: Generalized Random ForestsR package version 4.0.2.
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*(523), 1228–1242.
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.
Zhao, Z. (2008). Sensitivity of propensity score methods to the specifications. *Economics Letters, 98*, 309–319.

Authors and Affiliations

Hugo Bodory[1] · Martin Huber[2] · Michael Lechner[3,4,5,6,7]

✉ Martin Huber
martin.huber@unifr.ch

Hugo Bodory
hugo.bodory@unisg.ch

Michael Lechner
michael.lechner@unisg.ch

[1] Vice-President's Board (Research & Faculty), University of St. Gallen, Varnbüelstrasse 14, 9000 St. Gallen, Switzerland

[2] Department of Economics, University of Fribourg, Bd. de Pérolles 90, 1700 Fribourg, Switzerland

[3] Department of Economics, University of St. Gallen, Varnbüelstrasse 14, 9000 St. Gallen, Switzerland

[4] CEPR and PSI, London, UK

[5] CESIfo, Munich, Germany

[6] IAB, Nuremberg, Germany

[7] IZA, Bonn, Germany