Double Machine Learning for Panel Data

Paul Clarke[1] **Annalivia Polselli**[2]

[1]Institute for Social and Economic Research (ISER), University of Essex

[2]Institute of Analytics and Data Science (IADS), University of Essex

MiSoC Advisory Board Meeting

November 15, 2023

Machine Learning in a nutshell

- Machine Learning (ML) algorithms are powerful **predictive** and **classification** tools in AI and Computer Science
  - e.g., penalized regression, tree-based approaches, neural networks
  - or text analysis, image analysis
- Why ML?
  - **Advantages**: complexity reduction, flexibility, model selection
  - **Disadvantages**: interpretability, depending on hyperparameter tuning, stopping rules (regularisation), cross-validation to avoid overfitting, computationally intensive

Example with CART

ML and the social sciences

- In social sciences, increasing attention to use ML potential for **causal inference**
  - e.g., in labour economics (Davis and Heller, 2017; Lechner, 2019; Knaus et al., 2022; Cengiz et al., 2022), health economics (Heiler and Knaus, 2021; Di Francesco, 2022), environmental economics (Klosin and Vilgalys, 2022; Stetter et al., 2022)
- We focus on the use of ML in more **traditional econometrics/statistics**
  - By using generic ML methods to learn models, but using OLS/GMM to retrieve causal effect (e.g., Belloni et al., 2016; Chernozhukov et al., 2018; Nie and Wager, 2021; Chernozhukov et al., 2022)
  - Not building/modifying learning algorithms (e.g. Athey and Imbens, 2016; Wager and Athey, 2018; Athey et al., 2019; Künzel et al., 2019; Lechner and Mareckova, 2022)
- We use ML to boost existing statistical estimation approaches for observational panel data
  - Specifically, to learn nuisance parameters of the confounders

The causal model for panel data

- We consider Robinson (1988)'s partially linear regression model for panel data

$$y_{it} = d_{it}\theta + g(\mathbf{x}_{it}) + \alpha_i + u_{it},$$

under $\mathbb{E}(\alpha_i | d_{it}, \mathbf{x}_{it}) \neq 0$ (i.e. fixed effects assumption)

- Need to control for confounders, but $g(\mathbf{x}_{it})$ unknown
  - What variables? How many?
  - Linear or nonlinear?
- Use ML tools (e.g., Lasso, trees, random forests) to get $\widehat{g}(\mathbf{x}_{it})$

The ML plug-in problem

- Estimating θ from

$$y_{it} = d_{it}\theta + \widehat{g}(\mathbf{x}_{it}) + \alpha_i + u_{it},$$

but *regularization* and/or *overfitting* bias leads to $\sqrt{n}(\widehat{\theta} - \theta) \not\to 0$ (i.e. standard asymptotics invalid)

- Double ML (DML) by Chernozhukov et al. (2018)
  1. Sample splitting
  2. Orthogonalisation of the estimating equations for θ

- DML in two-stages:
  1. Learn nuisance parameters from each data fold ▸ Stage1
  2. Solve the sample analogue of the moment condition wrt θ.
     Average over folds attenuates ML bias and ensures $\sqrt{n}(\widehat{\theta} - \theta) \to 0$.
     ▸ Stage2 ▸ Variance

The estimators for panel data

- Estimate θ consistently from

$$y_{it} = d_{it}\theta + g(\mathbf{x}_{it}) + \alpha_i + u_{it},$$

  - under $\mathbb{E}(\alpha_i | d_{it}, \mathbf{x}_{it}) \neq 0$ (i.e. fixed effects assumption)
  - $g(\mathbf{x}_{it})$ may be non-linear in $\mathbf{x}$

- We develop CRE, WG and FD estimators
  CRE | WG/FD-hybrid | WG/FD-approx

- These are based on Mundlak (1978)'s model for fixed effect
  - Requires model assumptions for α_i
  - Flexible way of predicting the treatment

Results

- Bias reduction with CART/RF/Lasso with extended dictionary even for very non-linear functions, unlike OLS and Lasso w/t dictionary ▸ Plots
- Tree-based approaches under-estimate SD of sampling distribution
  - Estimators not normally distributed ▸ Distributions
  - Sensitivity to hyperparameter tuning (Machlanski et al., 2023) ▸ Tuning
- Same behaviour when applying DML to observational panel data ▸ Application
- Overall, ML are powerfull tools to approximate complex functional forms, but beware learner mismatch ("Don't use sledgehammer to crack nut")

Summary and future research

- This setting is simple but raised many challenges for ML
  - Non-linearity of the nuisance parameters
  - Presence of fixed effects
- Within the panel data framework, move on to
  - Heterogeneous treatment effects following Nie and Wager (2021)
  - CATEs (e.g., interactive regression model by Chernozhukov et al., 2018)
  - Dynamic panel models (cf Semenova et al., 2023)
- Allow for any learner within the context of DML (not only Lasso)
- Apply the method to other empirical questions
  - Impact of minimum wage on employment, voting behaviour, mental health, etc.
  - Impact of maternal smoking on child's outcomes at birth
  - Other possible applications ...

Thank you for your attention!

References I

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

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

Belloni, A., Chernozhukov, V., Hansen, C., and Kozbur, D. (2016). Inference in high-dimensional panel models with an application to gun control. *Journal of Business & Economic Statistics*, 34(4):590–605.

Cengiz, D., Dube, A., Lindner, A., and Zentler-Munro, D. (2022). Seeing beyond the trees: Using machine learning to estimate the impact of minimum wages on labor market outcomes. *Journal of Labor Economics*, 40(S1):S203–S247.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68.

Chernozhukov, V., Newey, W. K., and Singh, R. (2022). Automatic debiased machine learning of causal and structural effects. *Econometrica*, 90(3):967–1027.

Davis, J. M. and Heller, S. B. (2017). Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review*, 107(5):546–550.

Di Francesco, R. (2022). Aggregation trees. *CEIS Working Paper*.

Fazio, A. and Reggiani, T. (2023). Minimum wage and tolerance for high incomes. *European Economic Review*, 155:104445.

References II

Heiler, P. and Knaus, M. C. (2021). Effect or treatment heterogeneity? policy evaluation with aggregated and disaggregated treatments. *arXiv preprint arXiv:2110.01427*.

Klosin, S. and Vilgalys, M. (2022). Estimating continuous treatment effects in panel data using machine learning with an agricultural application. *arXiv preprint arXiv:2207.08789*.

Knaus, M. C., Lechner, M., and Strittmatter, A. (2022). Heterogeneous employment effects of job search programs: A machine learning approach. *Journal of Human Resources*, 57(2):597–636.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165.

Lechner, M. (2019). Modified causal forests for estimating heterogeneous causal effects. *arXiv preprint arXiv:1812.09487*.

Lechner, M. and Mareckova, J. (2022). Modified causal forest. *arXiv preprint arXiv:2209.03744*.

Machlanski, D., Samothrakis, S., and Clarke, P. (2023). Hyperparameter tuning and model evaluation in causal effect estimation. *arXiv preprint arXiv:2303.01412*.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, pages 69–85.

Nie, X. and Wager, S. (2021). Qusi-oracle estimation of heterogeneous treatment effects. *Biometrika*, pages 299–319.

References III

Robinson, P. M. (1988). Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 931–954.

Semenova, V., Goldman, M., Chernozhukov, V., and Taddy, M. (2023). Inference on heterogeneous treatment effects in high-dimensional dynamic panels under weak dependence. *Quantitative Economics*, 14(2):471–510.

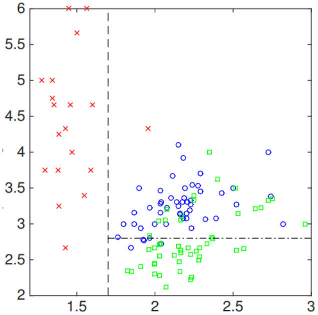
Stetter, C., Mennig, P., and Sauer, J. (2022). Using machine learning to identify heterogeneous impacts of agri-environment schemes in the eu: a case study. *European Review of Agricultural Economics*, 49(4):723–759.

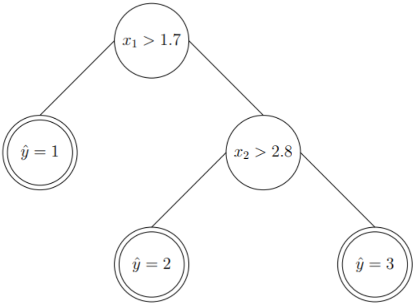
Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Yang, L., Liu, S., Tsoka, S., and Papageorgiou, L. G. (2017). A regression tree approach using mathematical programming. *Expert Systems with Applications*, 78:347–357.

CART ▸ Back

An simple illustration of a regression tree

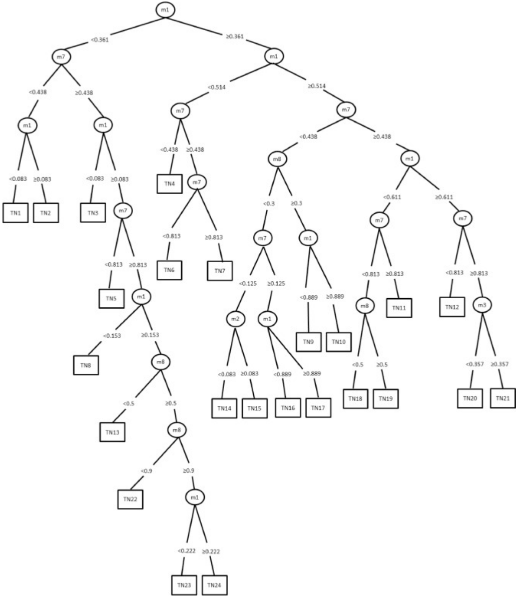




Source: https://wei2624.github.io/MachineLearning/sv_trees/

CART

▸ Back



Source: Yang et al. (2017). 'A regression tree approach using mathematical programming'.

Hyperparameter tuning ▸ Back

Learner	Hyperparamters	Value/interval	Description
Lasso	lambda.min	–	λ equivalent to minimum mean cross-validated error
CART	cp	{0.01,0.02}	Prune all nodes with a complexity less than cp from the printout.
	minbucket	{5,ceiling(N/2)}	Minimum number of observations in any terminal leaf node.
	minsplit	minbucket$\times$3	Minimal node size to split at.
	maxdepth	{1,10}	Maximum depth of any node of the final tree.
RF	num.trees	{5,100}	Number of trees in the forest.
	min.node.size	{5,ceiling(N/2)}	Minimal node size to split at.
	max.depth	{1,10}	Maximum depth of any node of the final tree.
	mtry	p	The number of covariates, randomly sampled, to split at each node.
	importance	impurity	The 'impurity' measure is the Gini index for classification, the variance of the responses for regression and the sum of test statistics.

Note: Hyperparameter tuning for CART and RF is conducted with a random grid search. For RF, nodes with size smaller than min.node.size can occur.

CRE or Mundlak's Device

Back

- Mundlak (1978) assumes $\mathbb{E}(\alpha_i|\mathbf{w}_{it}) = \overline{\mathbf{w}}_i\boldsymbol{\beta}$ where $\overline{\mathbf{w}}_i = T^{-1}\sum_{t=1}^{T}\mathbf{w}_{it}$
- Only works for nonlinear case if $l(.)$ known (i.e. oracle)
- Instead, we show equivalent learning problem is

$$y_{it} = v_{it}\theta + l_2(\mathbf{x}_{it}, \overline{\mathbf{x}}_i) + r_i + u_{it} \tag{1}$$

$$v_{it} = d_{it} - m_2(\mathbf{x}_{it}, \overline{\mathbf{x}}_i, \overline{d}_i), \tag{2}$$

  provided

  - $\alpha_i = \overline{\mathbf{x}}_i'\boldsymbol{\delta} + r_i$
  - Additively separable $d_{it} = m_1(\mathbf{x}_{it}, \overline{\mathbf{x}}_i) + c_i + v_{it}$
    $\implies m_2(\mathbf{x}_{it}, \overline{\mathbf{x}}_i, \overline{d}_i) = m_1(\mathbf{x}_{it}, \overline{\mathbf{x}}_i) + \overline{d}_i\pi - \overline{m}_1(\overline{\mathbf{x}}_i)$

WG and FD Transformations ▸ Back

- **Within-group (WG):** $\widetilde{w}_{it} = w_{it} - T^{-1} \sum_{s=1}^{T} w_{is}$

$$\widetilde{y}_{it} = \widetilde{d}_{it}\theta + \widetilde{l}(\mathbf{x}_{it}) + \widetilde{u}_{it} \tag{3}$$

$$\widetilde{v}_{it} = \widetilde{d}_{it} - \widetilde{m}(\mathbf{x}_{it}) \tag{4}$$

$\widetilde{l}(\mathbf{x}_{it}) \equiv l(\mathbf{x}_{it}) - T^{-1} \sum_{s=1}^{T} l(\mathbf{x}_{is}) \approx l(\widetilde{\mathbf{x}}_{it})$
$\widetilde{m}(\mathbf{x}_{it}) \equiv m(\mathbf{x}_{it}) - T^{-1} \sum_{s=1}^{T} m(\mathbf{x}_{is}) \approx m(\widetilde{\mathbf{x}}_{it})$

- **First-difference (FD):** $\Delta w_{it} = w_{it} - w_{it-1}$

$$\Delta y_{it} = \Delta d_{it}\theta + \Delta l(\mathbf{x}_{it}) + \Delta u_{it} \tag{5}$$

$$\Delta v_{it} = \Delta d_{it} - \Delta m(\mathbf{x}_{it}) \tag{6}$$

for $t = 2, \ldots, T$

$\Delta l(\mathbf{x}_{it}) \equiv l(\mathbf{x}_{it}) - l(\mathbf{x}_{it-1}) \approx l(\Delta \mathbf{x}_{it})$
$\Delta m(\mathbf{x}_{it}) \equiv m(\mathbf{x}_{it}) - m(\mathbf{x}_{it-1}) \approx m(\Delta \mathbf{x}_{it})$

Hybrid alternative for WG and FD estimators ▸ Back

1. Learn $l(\mathbf{x}_{it})$ and $m(\mathbf{x}_{it})$ using Mundlak (1978)'s device for CRE
2. Apply WG or FD transformation to $\widehat{l}(\mathbf{x}_{it})$ and $\widehat{m}(\mathbf{x}_{it})$
   - WG transformation:
     $\widetilde{\widehat{l(\mathbf{x}_{it})}} = \widehat{l}(\mathbf{x}_{it}) - T^{-1}\sum_{s=1}^{T}\widehat{l}(\mathbf{x}_{is})$
     $\widetilde{\widehat{m(\mathbf{x}_{it})}} = \widehat{m}(\mathbf{x}_{it}) - T^{-1}\sum_{s=1}^{T}\widehat{m}(\mathbf{x}_{is})$
   - FD transformation:
     $\Delta\widehat{l(\mathbf{x}_{it})} = \widehat{l}(\mathbf{x}_{it}) - \widehat{l}(\mathbf{x}_{it-1})$
     $\Delta\widehat{m(\mathbf{x}_{it})} = \widehat{m}(\mathbf{x}_{it}) - \widehat{m}(\mathbf{x}_{it-1})$

Estimating θ with DML: Stage 1 ▸ Back

- Split the sample $\mathcal{W}$ in k folds s.t.
  - Unit $i \in \mathcal{W}_k \Rightarrow i \notin \mathcal{W}_k^c$
  - Unit i is observed for T_i time periods
- Learn the nuisance parameters $\eta = (l, m)$ for each fold using $i \notin \mathcal{W}_k$
- Plug $\widehat{\eta}_k$ in the orthogonal score for the k-th fold

$$\psi(\widetilde{W}_{it};\theta,\eta_k) = \big(\widetilde{d}_{it} - \widetilde{m}(.)_k\big)\Big\{\big(\widetilde{y}_{it} - \widetilde{l}(.)_k - \big(\widetilde{d}_{it} - \widetilde{m}(.)_k\big)\theta\Big\} \tag{7}$$

for all $i \in \mathcal{W}_k$ and $t \in S_i$, where $\widetilde{m}(.) = \mathbb{E}\big(\widetilde{d}_{it}|\widetilde{\mathbf{x}}_{it}\big)$ and $\widetilde{l}(.) = \mathbb{E}\big(\widetilde{y}_{it}|\widetilde{\mathbf{x}}_{it}\big)$

Estimating θ with DML: Stage 2 Back

- By Chernozhukov et al. (2018)'s Lemma (2.6), θ satisfies

$$\mathbb{E}\left[\psi(\widetilde{W};\theta,\eta)\right] = 0 \tag{8}$$

- $\widehat{\theta}$ solves the finite-sample analog of (8) for each fold k

$$\frac{1}{|\mathcal{W}_k|}\sum_{i\in\mathcal{W}_k}\sum_{t\in S_i}\psi(\widetilde{W}_{it};\theta,\widehat{\eta}_k) = 0 \tag{9}$$

  with closed-form solution

$$\widehat{\theta}_k = \left(\frac{1}{|\mathcal{W}_k|}\sum_{i\in\mathcal{W}_k}\sum_{t\in S_i}\widehat{v}_{it}^2\right)^{-1}\frac{1}{|\mathcal{W}_k|}\sum_{i\in\mathcal{W}_k}\sum_{t\in S_i}\widehat{v}_{it}\left(\widetilde{y}_{it} - \widehat{\widetilde{l}(.)}_k\right) \tag{10}$$

  where $\widehat{v}_{it} = \widetilde{d}_{it} - \widehat{\widetilde{m}(.)}_k$.

- Average over all k folds to get DML estimator of θ

Inference with DML

Back

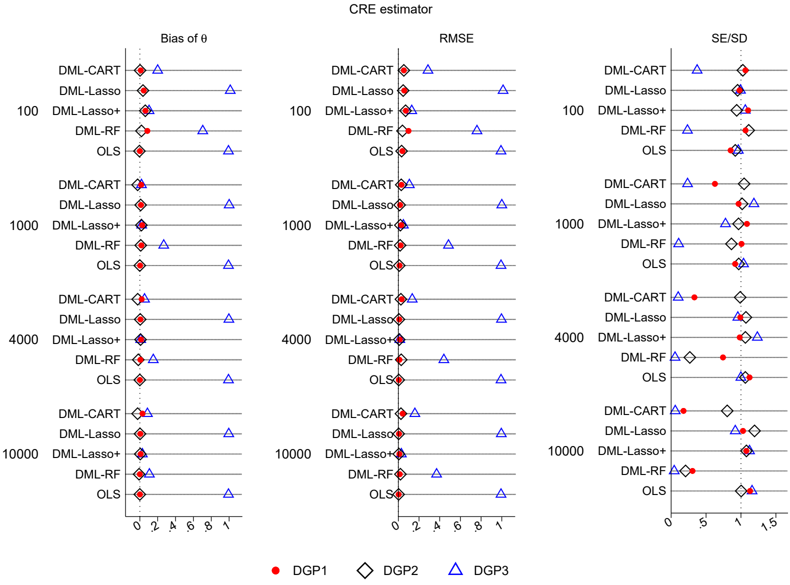
- The estimator of the cluster-robust variance is

$$\widehat{\sigma}^2 = \widehat{J}_0^{-1} \left[\frac{1}{|\mathcal{W}_k|} \sum_{i \in \mathcal{W}_k}^{N} \sum_{t \in S_i} \psi(\widetilde{W}_{it}; \theta, \widehat{\eta}_k) \psi(\widetilde{W}_{it}; \theta, \widehat{\eta}_k)' \right] \widehat{J}_0^{-1}$$

where $\widehat{J}_0 = \frac{1}{|\mathcal{W}_k|} \sum_{i \in \mathcal{W}_k}^{N} \sum_{t \in S_i} \left(\widetilde{d}_{it} - \widetilde{m}\widehat{(.)} \right)^2 \ \forall\, k = 1, \ldots, K$

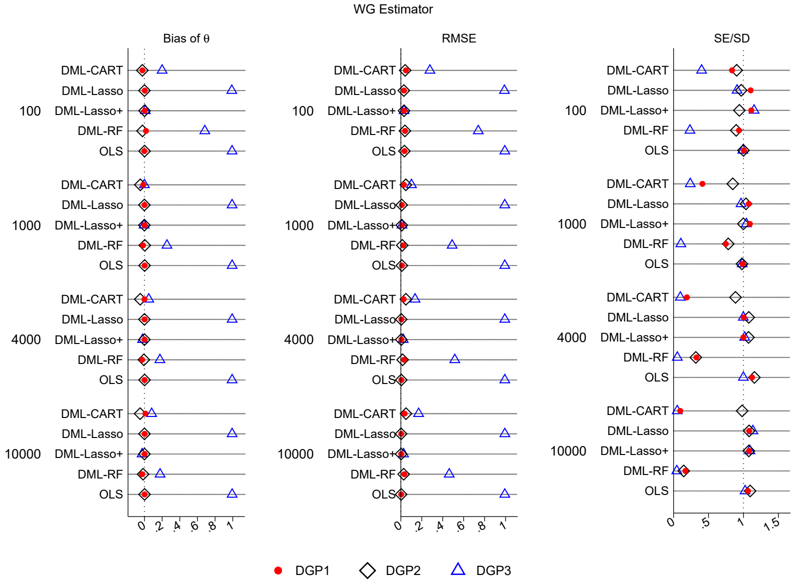
MC Simulation results

Back



Note: Averages over 100 Monte Carlo replications. DGP1 is linear in the nuisance functions; DGP2 smooth non-linear; DGP3 non-smooth non-linear. Time if fixed to $t = 10$ time periods.

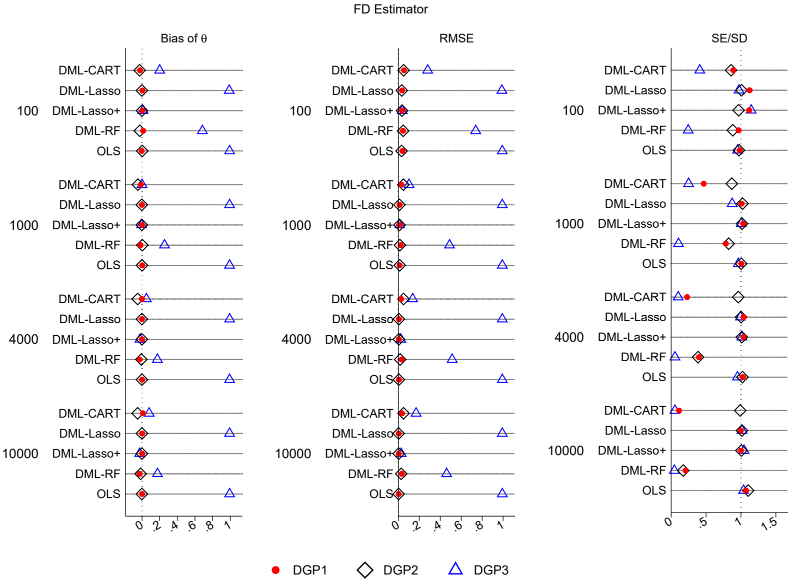
MC Simulation results ▸ Back



Note: Averages over 100 Monte Carlo replications. DGP1 is linear in the nuisance functions; DGP2 smooth non-linear; DGP3 non-smooth non-linear. Time if fixed to $t = 10$ *time periods.*

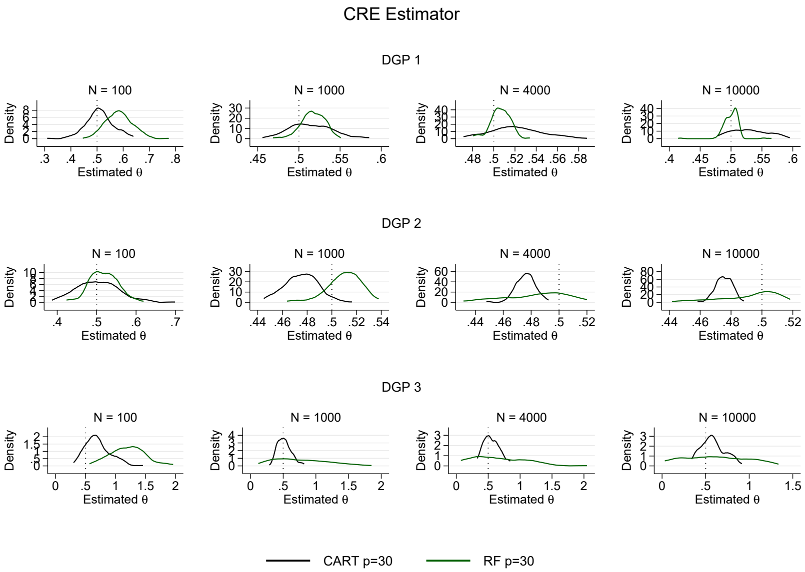
MC Simulation results

Back



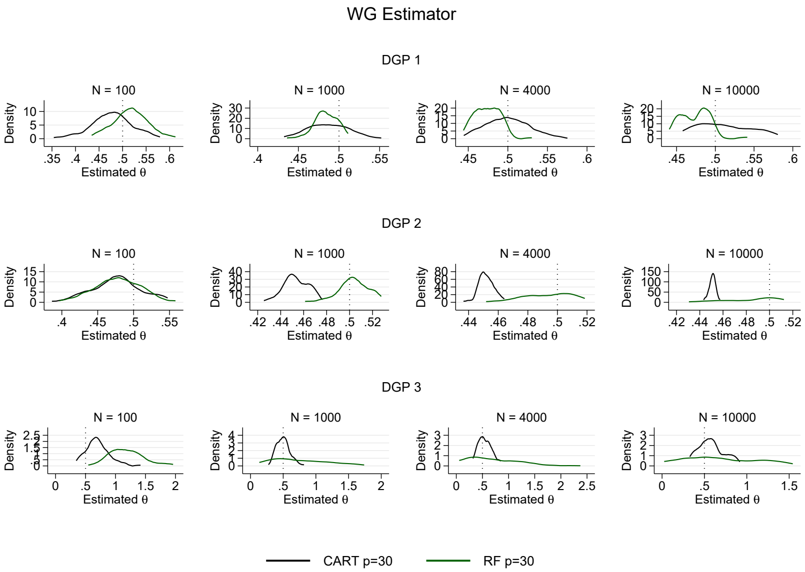
Note: Averages over 100 Monte Carlo replications. DGP1 is linear in the nuisance functions; DGP2 smooth non-linear; DGP3 non-smooth non-linear. Time if fixed to $t = 10$ time periods.

Sampling distribution of $\widehat{\theta}$ ▸ Back



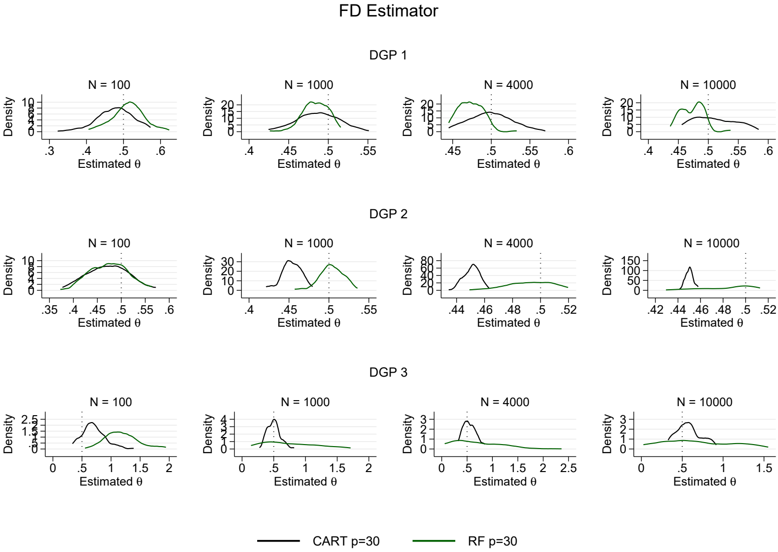
Note: The total number of variables is twice p because individual means are included as inputs following Mundlak (1978).

Sampling distribution of $\widehat{\theta}$ Back



Note: The total number of variables is twice p because individual means are included as inputs following Mundlak (1978).

Sampling distribution of $\widehat{\theta}$ Back



Note: The total number of variables is twice p because individual means are included as inputs following Mundlak (1978).

Application with National Minimum Wage ▸ Back

- We re-asses ‘Minimum wage and tolerance for high incomes’ by Fazio and Reggiani (2023, EER) using DML for panel data
- We replicate Specification (2) in Table (5)
  - Investigating voting behaviour after the introduction of the National Minimum Wage (NMW) in the UK in 1999.
  - They find that having benefited from the NMW raises the probability of voting conservative parties.
- Data: British Household Panel Survey (BHPS)

	OLS (1)	OLS (2)	DML-Lasso (3)	DML-Lasso (4)	DML-CART (5)	DML-RF (6)
Dependent variable: "Vote conservative"						
NMW	0.097**	0.088**	0.093***	0.079**	0.089**	0.149
	(0.045)	(0.045)	(0.045)	(0.045)	(0.05)	(0.127)
No. Observations	19,961	19,961	19,961	19,961	19,961	19,961
No. Groups	4,927	4,927	4,927	4,927	4,927	4,927
Controls vars	Yes	Yes	Yes	Yes	Yes	Yes
Wave FE	Yes	Yes	Yes	Yes	Yes	Yes
Region FE	Yes	Yes	Yes	Yes	Yes	Yes
Wave x Region FE	No	Yes	No	Yes	No	No
Resampling Information						
Estimator	WG	WG	WG-hybrid	WGG-hybrid	WGG-hybrid	WGG-hybrid
No. folds	–	–	5	5	5	5
Folds per cluster	–	–	5	5	5	5
No. repeated sample splits	–	–	1	1	1	1
Cross-fitting	–	–	Yes	Yes	Yes	Yes
Score	–	–	PO	PO	PO	PO
DML algorithm	–	–	2	2	2	2

*Note: Column (1) reports the figures of Specification (2) in Table 5 in Fazio and Reggiani (2023) estimated using least squares; Column (2) is LS regression with controls from Column (1) and interaction terms between wave and region fixed effects; remaining columns use DML with different (tuned) learners. Control variables include: age, education, marital status, household size, income of other members, and their individual means. Age squared and is included in Columns (1)-(3). Standard errors (in parenthesis) are clustered at the individual level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.*

Back