Double/Debiased Machine Learning comparison study

November 9, 2022

Abstract

TBA

1 Motivation

Double/Debiased machine learning model is especially suited for the analysis of causal effects.

2 Double/Debiased Machine Learning Model

2.1 Partial linear regression model

We want to analyze the relationship of an $(n \times 1)$ outcome variable Y and a $(n \times 1)$ treatment or policy variable D. There is a $(n \times k)$ matrix X of k $(n \times 1)$ variables x_l, $l = 1, ..., k$ which are related to Y but are not of primary interest. These variables are related to the treatment D and are therefore denoted as confunding or control variables. We assume the following partially linear regression (PLR) model for the relationship of Y, D and X [5].

$$Y = D\theta_0 + g_0(X) + U, \quad E[U|X, D] = 0 \tag{1}$$

$$D = m_0(X) + V, \quad E[V|X] = 0 \tag{2}$$

g_0 and m_0 are function of mapping X to $\mathbb{R}$. g_0 and m_0 are also denoted as nuisance functions because their specific functional forms are not of central interest. U and V are $(n \times 1)$ vectors of error terms or disturbances. θ_0 is scalar parameter that measures the effect of D on Y and is of central interest.[1] $E[V|X]$ denotes the expected value of the variable V given the variables X.

Note that we cannot estimate g_0 from Equation (1) alone. The reason is that not only g_0 is a function of X but also D is via (2). Denote the approach where we estimate g_0 from Equation (1) alone as naive approach. The in the naive approach, we would for example iterate until convergence between updating g_0 using $Y - D\theta_0 = g_0(X) + U$ and updating θ_0 using $Y - g_0(X) = D\theta_0 + U$. An estimate of g_0 from the naive approach would be biased as it does not account for the fact that $Y - D\theta_0$ is depended of X, i.e., $m_0(X) = E[X|D]$. However $g_0(X) \neq E[Y|X]$ because Y depends also on $D\theta_0$.

[5] propose an algorithm to estimate the treatment effect parameter θ_0 from the PLR model from above unbiased and $n^{-1/2}$ root consistent using machine learning (ML) to assess g_0 and m_0 if the ML estimators are at least $n^{-1/4}$ root consistent. This is where the name Double/Debiased ML (DML) comes from. The algorithm uses cross fitting and the idea of partialing-out to avoid the regularization bias. We outline the latter in more detail below.

1. Estimate $D = \hat{m}_0(X) + \hat{V}$. This is the treatment model where we now partial out the effect of X from D.
2. Estimate $Y = D\hat{\theta}_0 + \hat{g}_0(X) + \hat{U}$ as in the naive approach from above. This is the outcome model where we with $Y - \hat{g}_0(X)$ partial out the effect of X on Y conditional on D.
3. Regress $Y - \hat{g}_0(X)$ on $\hat{V}$ using ordinary least squares (OLS) and denote the resulting parameter as $\check{\theta}_0$.

[1] Note that we assume that the treatment D is additive linear related to Y. Doing so helps for an easily understandable description of the approach. We will relax the assumption of linear additivity in next subsection below.

$\check{\theta}_0$ is treatment effect estimate that is free of regularization bias. Note that the partialing-out approach is similar to the residuals-on-residuals regression for linear regression models [7, 8, 9] and for kernel regression models [10]. We could obtain $\check{\theta}_0$ also if we replaced Step 2. from the partialing-out approach from above with the following $Y = \hat{l}_0(X) + \hat{U}$ and regress in Step 3. $\hat{U}$ on $\hat{V}$ using OLS. l_0 is a function that maps X to $\mathbb{R}$ reflects $E[Y|X]$.

2.2 PLR model modifications

We can extend the PLR model to the following partial linear regression instrumental variables model (PLR-IV).

$$\begin{aligned} Y &= D\theta_0 + g_0(X) + U, \quad E[U|X, Z] = 0 \\ Z &= m_0(X) + V, \quad E[V|X] = 0. \end{aligned}$$

Z is $n \times l$ matrix of instrumental variables. The usage of the instrumental variables allows us to account for the endogeneity of the treatment variables D, i.e., that D is (indirectly) correlated with U. This is because the error term U is correlated with V and V is correlated with D. We can combine Equations (3) and (3) to build score function for the inference of θ and the nuisance functions. See [5] for more details.

The PLR model can be generalized the following interactive regression model (IRM):

$$\begin{aligned} Y &= g_0(X, D) + U, \quad E[U|X, D] = 0 \\ D &= m_0(X) + V, \quad E[V|X] = 0. \end{aligned}$$

Now the function g_0 from Equation (3) jointly maps X and D to $\mathbb{R}$. Typically, the treatment variable D is now constrained to be binary, with values 0 and 1. This way, the average treatment effect (ATE) (and the local average treatment effect, LATE) can be estimated as in [5].

The IRM model can be further generalized to an interactive regression model using instrumental variables (IIV) as follows:

$$\begin{aligned} Y &= \mu_0(X, Z) + U, \quad E[U|X, Z] = 0 \\ D &= m_0(X, Z) + V, \quad E[V|X, Z] = 0 \\ Z &= p_0(X) + \zeta, \quad E[\zeta|X] = 0. \end{aligned}$$

μ_0 is a function that maps X and Z to $\mathbb{R}$.

3 Simulation experiments

We conduct in total four simulation experiments, i.e., Scenario 1-4. The data generating processes (DGPs) increase in terms of complexity for the scenarios and we compare the results from models which are suitable form the underlying data structure. In each simulation experiment, we create $n = 1.000$ observations of data for $k = 20$ variables. We set the treatment effect parameter θ to 0.5 and replicate the DGP 30 times. The DGPs of the simulation experiments (A)-(D) are from [1] and [2].

To estimate the non-linear functions in the PLR, IRM, PLR-IV and IIV models, we use random forest models, either classification models if the dependent variable is binary and regression model otherwise. As hyper-parameters, we sue 100 trees, a maximum of 20 features, a maximum depth of 5 and a minimum sample of 2 per leave.

In addition to the PLR, IRM, PLR-IV and IIV models, we compare the results of the following three models (I)-(III) when suitable. (I) An linear regression model, denoted as OLS of the following form.

$$Y = D\theta + X\beta + U.$$

(II) A two stage least squares regression model, denoted as 2SLS of the following form.

$$\begin{aligned} Y &= D\theta + X\beta sls + U, \\ W &= X\gamma + Z\delta + \zeta + V, \end{aligned}$$

where W is a $(n \times (k+1))$ matrix that consists of the matrix X with an additional column which is D. (III) A naive ML model, denoted as naive-ML of the following form.

$$Y = t_0(D, X) + U$$

where t_0 is a function that maps D and X to $\mathbb{R}$.

3.1 Scenario 1

3.1.1 DGP: PLR model

The Scenario 1 represents the case of a PLR model and we use the following DGP for $i = 1, ..., n$ [5].

$$\begin{aligned} y_i &= \theta d_i + g_0(x_i) + s_2\zeta_i, \\ d_i &= m_0(x_i) + s_1 v_i, \\ x_i &\sim \mathcal{N}(0, \Sigma), \\ \zeta_i &\sim \mathcal{N}(0, 1), \\ v_i &\sim \mathcal{N}(0, 1), \end{aligned}$$

where d_i, v_i, y_i and ζ_i are the i^{th} entries of D, V, Y and ζ, respectively. $\mathcal{N}(\mu_n, \Sigma_n)$ represents the normal distribution with mean value μ_n and variance Σ_n. Note that μ_n can be a vector and Σ_n is in this case a variance-covariance matrix. x_i is the $(k \times 1)$ vector of row i from matrix X. Σ is a matrix with entries $\sigma_{mj} = 0.7^{|j-m|}$, with $m = 1, .., k$ and $j = 1, ..., k$. The nuisance functions are given by

$$\begin{aligned} m_0(x_i) &= a_0 x_{i,1} + a_1 \frac{\exp(x_{i,3})}{1 + \exp(x_{i,3})}, \\ g_0(x_i) &= b_0 \frac{\exp(x_{i,1})}{1 + \exp(x_{i,1})} + b_1 x_{i,3}. \end{aligned}$$

We use the following parameter values: $a_0 = 1, a_1 = 0.25, s_1 = 1, b_0 = 1, b_1 = 0.25$ and $s_2 = 1$. Note that the nuisance functions m_0 and g_0 are non-linear in x_i.

3.1.2 Results

The Figure 1 shows the true and estimated treatment effects per replication for Scenario 1. The error bars represent the upper and lower bound of the 95% confidence interval of the estimate. The red line represents the true value of the treatment effect.

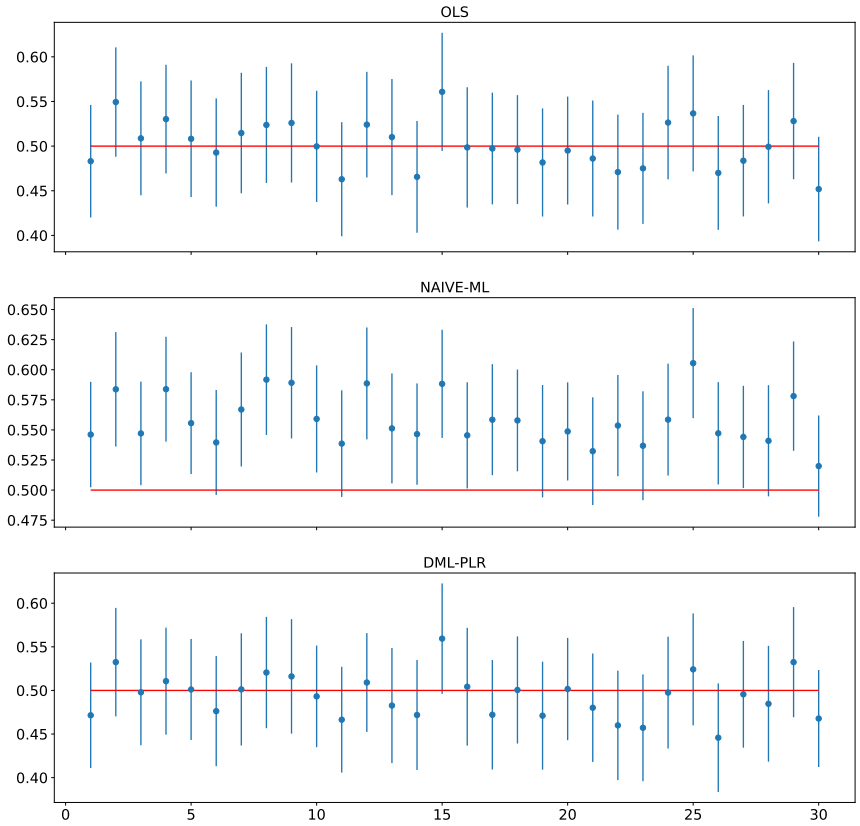

Figure 1: True and estimated treatment effects per replication for Scenario 1. The error bars represent the upper and lower bound of the 95% confidence interval of the estimate. The red line represents the true value of the treatment effect.

We find that the DML-PLR and the OLS model tend to recover the treatment effect overall similarly in terms of how accurate the point estimates are and of how wider the confidence intervals are. The NAIVE-ML model has the tendency to overfitted the treatment effect.

Table 1 shows the root mean squared error (RMSE), the mean absolute error (MAE) and the bias of estimated treatment effect and the true value across the replications for the compared models. The last row indicates which model performs best according to RMSE, MAE or bias.

	RMSE	MAE	Bias
OLS	0.0262	0.0212	0.0019
NAIVE-ML	0.0617	0.0582	0.0582
DML-PLR	0.0261	0.0207	-0.0065
Best	DML-PLR	DML-PLR	OLS

Table 1: Root mean squared error (RMSE), mean absolute error (MAE) and bias of estimated treatment effect and the true value across the replications for the compared models. The last row indicates which model performs best according to RMSE, MAE or bias.

3.2 Scenario 2

3.2.1 DGP: IRM model

The Scenario 2 represents the case of a IRM model and we use the following DGP for $i = 1, ..., n$. The data generating process is based on a the simulation experiment in [3].

$$\begin{aligned} y_i &= \theta d_i + c_y x_i' \beta d_i + \zeta_i, \\ d_i &= 1\left\{\frac{\exp(c_d x_i'\beta)}{1+\exp(c_d x_i'\beta)} > v_i\right\}, \\ \zeta_i &\sim \mathcal{N}(0,1), \\ v_i &\sim \mathcal{U}(0,1), \\ x_i &\sim \mathcal{N}(0,\Sigma), \end{aligned}$$

where Σ is a matrix with entries $\Sigma_{kj} = 0.5^{|j-m|}$, with $m = 1, .., k$ and $j = 1, ..., k$. $\mathcal{U}(a, b)$ represents the continuous uniform distribution with parameters a and b. β is a $(k \times 1)$ vector with entries $\beta_j = \frac{1}{j^2}$ and the constants c_y and c_d are the following:

$$c_y = \sqrt{\frac{R_y^2}{(1-R_y^2)\beta'\Sigma\beta}}, \qquad c_d = \sqrt{\frac{(\pi^2/3)R_d^2}{(1-R_d^2)\beta'\Sigma\beta}}.$$

We set the parameters R_d^2 and R_y^2 to 0.5.

3.2.2 Results

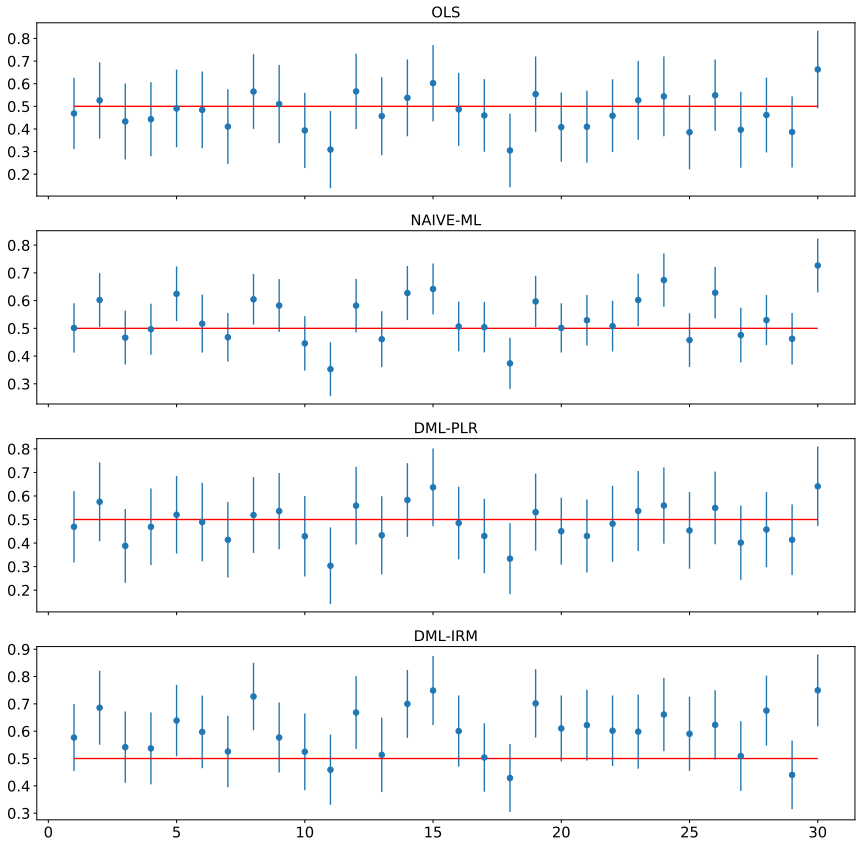

Figure 2: True and estimated treatment effects per replication for Scenario 2. The error bars represent the upper and lower bound of the 95% confidence interval of the estimate. The red line represents the true value of the treatment effect.

	RMSE	MAE	Bias
OLS	0.0854	0.0699	-0.0269
NAIVE-ML	0.0922	0.0710	0.0350
DML-PLR	0.0809	0.0671	-0.0174
DML-IRM	0.1306	0.1094	0.0979
Best	DML-PLR	DML-PLR	DML-PLR

Table 2: Root mean squared error (RMSE), mean absolute error (MAE) and bias of estimated treatment effect and the true value across the replications for the compared models. The last row indicates which model performs best according to RMSE, MAE or bias.

3.3 Scenario 3

3.3.1 PLR-IV model

The Scenario 3 represents the case of a PLR-IV model and we use the following DGP for $i = 1, ..., n$ [4].

$$\begin{aligned} y_i &= \theta d_i + x_i'\beta + \varepsilon_i, \\ d_i &= x_i'\gamma + z_i'\delta + u_i, \\ z_i &= \Pi x_i + \zeta_i, \end{aligned}$$

with

$$\begin{pmatrix} \varepsilon_i \\ u_i \\ \zeta_i \\ x_i \end{pmatrix} \sim \mathcal{N}\left(0, \begin{pmatrix} 1 & 0.6 & 0 & 0 \\ 0.6 & 1 & 0 & 0 \\ 0 & 0 & 0.25 I_l & 0 \\ 0 & 0 & 0 & \Sigma \end{pmatrix}\right)$$

where Σ is a $k \times k$ matrix with entries $\Sigma_{mj} = 0.5^{|j-m|}$, with $m = 1, .., k$ and $j = 1, ..., k$. I_l is the $l \times l$ identity matrix. $\beta = \gamma$ is a k-vector with entries $\beta_j = \frac{1}{j^2}$ with $j = 1, ..., k$. δ is a l-vector with entries $\delta_j = \frac{1}{h^2}$, with $h = 1, ..., l$. Π is a matrix of parameters and specified as follows: $\Pi = (I_l, 0_{l \times (k-l)})$, where $0_{l \times (k-l)}$ is a $(l \times (k-l))$ matrix of zeros. Note that the endogeneity of D comes from the non-zero correlation of ε_i and u_i.

3.3.2 Results

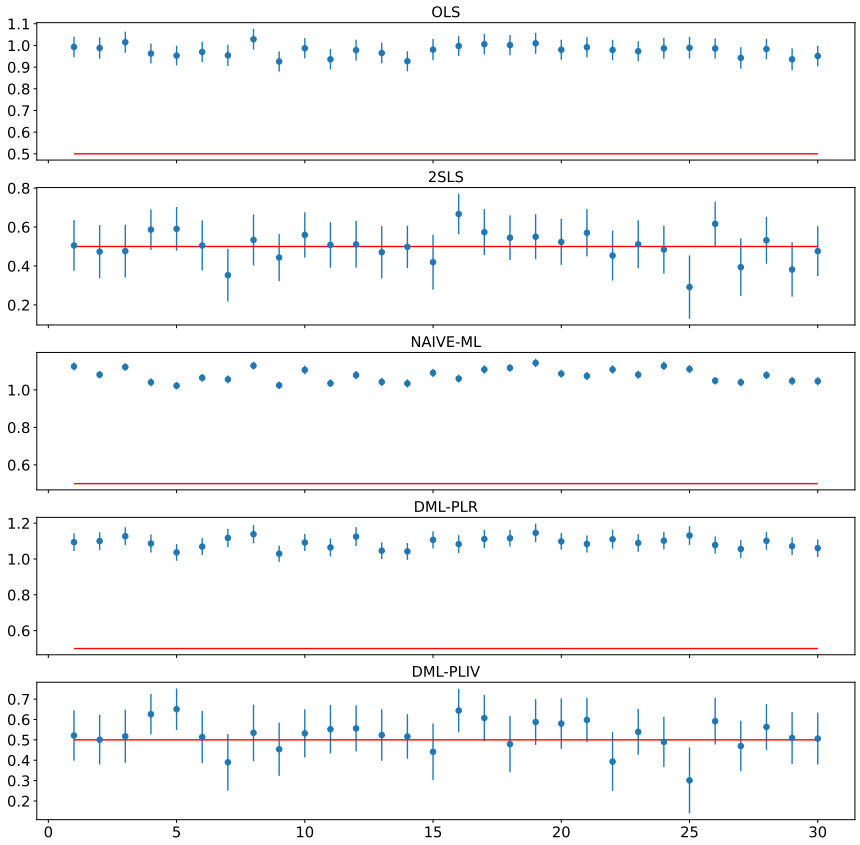

Figure 3: True and estimated treatment effects per replication for Scenario 3. The error bars represent the upper and lower bound of the 95% confidence interval of the estimate. The red line represents the true value of the treatment effect.

	RMSE	MAE	Bias
OLS	0.4765	0.4758	0.4758
2SLS	0.0783	0.0591	0.0001
NAIVE-ML	0.5788	0.5778	0.5778
DML-PLR	0.5912	0.5904	0.5904
DML-PLIV	0.0792	0.0618	0.0230
Best	2SLS	2SLS	2SLS

Table 3: Root mean squared error (RMSE), mean absolute error (MAE) and bias of estimated treatment effect and the true value across the replications for the compared models. The last row indicates which model performs best according to RMSE, MAE or bias.

3.4 Scenario 4

3.4.1 IIV model

The Scenario 4 represents the case of a IIV model and we use the following DGP for $i = 1, ..., n$. The DGP is based on the simulation experiment of [6].

$$\begin{aligned} y_i &= \theta d_i + x_i'\beta + u_i, \\ d_i &= 1\{\alpha_x Z + v_i > 0\}, \end{aligned}$$

and

$$\begin{aligned}\begin{pmatrix} u_i \\ v_i \end{pmatrix} &\sim \mathcal{N}\left(0, \begin{pmatrix} 1 & 0.3 \\ 0.3 & 1 \end{pmatrix}\right), \\ Z &\sim \text{Bernoulli}(0.5), \\ x_i &\sim \mathcal{N}(0, \Sigma),\end{aligned}$$

where Bernoulli(p) represents the Bernoulli distribution with parameter p. Σ is a matrix with entries $\Sigma_{kj} = 0.5^{|j-m|}$, with $m = 1, .., k$ and $j = 1, ..., k$. β is a $(k \times 1)$ vector with entries $\beta_j = \frac{1}{j^2}$ for $j = 1, ..., k$ and we set α_x to one. Note that the endogeneity of D comes from the non-zero correlation of u_i and v_i.

3.4.2 Results

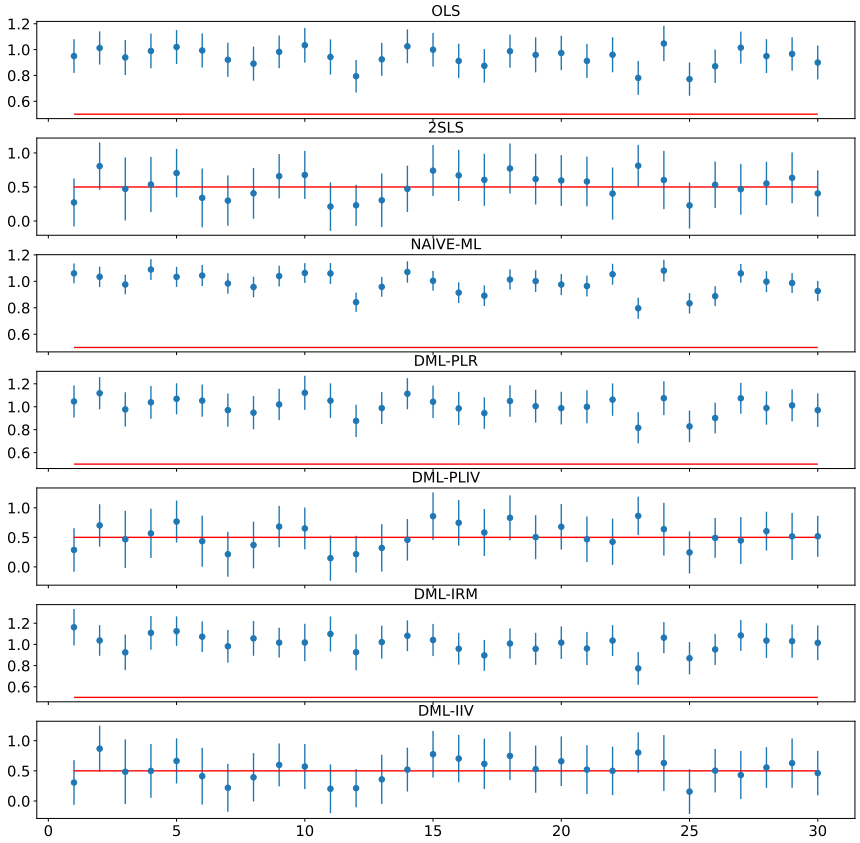

Figure 4: True and estimated treatment effects per replication for Scenario 4. The error bars represent the upper and lower bound of the 95% confidence interval of the estimate. The red line represents the true value of the treatment effect.

	RMSE	MAE	Bias
OLS	0.4489	0.4433	0.4433
2SLS	0.1768	0.1527	0.0204
NAIVE-ML	0.4926	0.4867	0.4867
DML-PLR	0.5103	0.5047	0.5047
DML-PLIV	0.1946	0.1575	0.0233
DML-IRM	0.5176	0.5114	0.5114
DML-IIV	0.1805	0.1420	0.0177
Best	2SLS	DML-IIV	DML-IIV

Table 4: Root mean squared error (RMSE), mean absolute error (MAE) and bias of estimated treatment effect and the true value across the replications for the compared models. The last row indicates which model performs best according to RMSE, MAE or bias.

4 Conclusion

Appendix

References

[1] Bach, P., Chernozhukov, V., Kurz, M. S., and Spindler, M., *DoubleML - An Object-Oriented Implementation of Double Machine Learning in Python, Journal of Machine Learning Research*, 23(53): 1-6, 2022.

[2] Bach, P., Chernozhukov, V., Kurz, M. S., and Spindler, M., *DoubleML - An Object-Oriented Implementation of Double Machine Learning in R*, 2021.

[3] Belloni, A., Chernozhukov, V., Fernández-Val, I. and Hansen, C., *Program Evaluation and Causal Inference With High-Dimensional Data*, Econometrica, 85: 233-298, 2017.

[4] Chernozhukov, V., Hansen, C. and Spindler, M., *Post-Selection and Post-Regularization Inference in Linear Models with Many Controls and Instruments*, American Economic Review: Papers and Proceedings, 105 (5): 486-90, 2015.

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

[6] Farbmacher, H., Guber, R. and Klaaßen, S., *Instrument Validity Tests with Causal Forests*, MEA Discussion Paper, 13-2020, 2020.

[7] Frisch R., Waugh F., *Partial Time Regressions as Compared with Individual Trends*, Econometrica, 1 (4), 387–401, 1933.

[8] Lovell M., *Seasonal Adjustment of Economic Time Series and Multiple Regression Analysis*, Journal of the American Statistical Association, 58 (304), 993–1010, 1963. .

[9] Lovell M., *A Simple Proof of the FWL Theorem*, Journal of Economic Education, 39 (1), 88–91, 2008.

[10] Robinson P., *Root-N-Consistent Semiparametric Regression*, Econometrica, 56(4):931-954, 1988.