

Is Double Machine Learning always better than Simple Linear Regression to estimate Causal Effects: Evidence from four simulation experiments

German Zenetti

April 4, 2023

Abstract

The double/debiased machine learning (DML) model is increasing in popularity as it is designed to use machine learning for the analysis of causal effects. We compare the performance of four variations of the DML model with a naive machine learning model and simple linear regression models. For doing so, we compare the estimation results for four different simulation experiment scenarios implemented in the Python package "DoubleML" from [1]. The simulation experiment scenarios are either with linear or non-linear effects and with or without endogeneity of the causal effect variable and base on widespread data generating processes for DML models in the literature and software implementation.

We find that the simple linear regression models perform similarly well overall as the DML models regarding, e.g., the bias or root mean squared error of the true and estimated causal effect. This outcome is at first sight surprising as the linear models do not account for the confounded causal effect variables and the non-linearity of the functional form in the data generating process. The reasons for the result are presumably that the DML models have with increasing complexity of the model to trade off their advantageous flexibility with a higher uncertainty of the underlying functional form and hence a decrease in the efficiency to recover the causal effect.

1 Motivation

Double/Debiased machine learning (DML) model is especially suited for the analysis of causal effects. Machine learning is often said to be good for prediction but bad for the inference of causal effects (see the next section for more details). This is where DML come into play as it is designed to analyze causal effects using a machine learning approach.

Causal effects are often a key outcome of statistical analysis. For example, a brand wants to know the effect of advertising investments on its sales performance. Or a manufacturer want to know the how strong the customer demand for his product decreases if he increase the price. In the best scenario, we would study these effect in a controlled experiments with many test units and various occasions (i.e., with many observations) where we control for all confounding variables and isolate the effect of interest. Ideally the set of observations who receive the treatment (e.g., are exposed to the advertisement) and who don't (e.g., are not exposed to the advertisement) are equal in terms of all other variables that are related to the outcome variable (e.g., their purchase decision for the advertised brand). This way would be able to estimate the causal effect of the treatment in a unbiased manner. However, often the scenario from a above is not possible or would be very expensive to realize in a typical real world setting.

In a typical real world setting, we may not have equal sets of observations who receive and don't receive the treatment, but whether an observations receives the treatment may depend on other variables. In the example from above, the brand may wish to advertise on e.g., online product pages with many customers visiting and not on others with few visitors. Or a brand may want to advertise more when the price is high and less when it is low. To deal with such as a real world setting, DML combines the power of machine learning and the effect of confounding variables to estimate the causal effect.

The following analysis bases on the Python package "DoubleML" from [1]. We will use the implementation and the data generation processes (DGP) for different variations of the DLM approach and compare their performance. All scripts for the analysis in this github repository: <https://github.com/g-r-m-n/dml>.

2 DML 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 , $l = 1, \dots, k$ which are related to Y but are not of primary interest. These variables are related to the treatment D and are therefore denoted as confounding or control variables. We assume the following partially linear regression (PLR) model for the relationship of Y , D and X [5].

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

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

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). To understand the underlying relationship, note that if m_0 maps X to zero or to a constant, D would be independent from X and we could identify θ_0 without considering Equation (2). Opposingly, the closer the variation of V is towards zero, the more D is explained by X alone and the larger the correlation of D and X .

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 causal 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 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 $\check{\theta}_0$.

$\check{\theta}_0$ is causal 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).

$$Y = D\theta_0 + g_0(X) + U, \quad E[U|X, Z] = 0$$

$$Z = m_0(X) + V, \quad E[V|X] = 0.$$

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):

$$Y = g_0(X, D) + U, \quad E[U|X, D] = 0$$

$$D = m_0(X) + V, \quad E[V|X] = 0.$$

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].

¹Note 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.

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

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

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 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 $n = 10,000$ observations of data for $k = 20$ variables. We set the causal effect parameter θ to 0.5 and replicate the DGP 100 times. The DGPs of the simulation experiments (A)-(D) are from [1] and [2].

Note that we find overall similar performance differences in the comparison of the different approaches in terms of that the double ML models (i.e., PLR, IRM, PLR-IV and IIV) are typically in a similar ballpark as their linear counterparts (i.e., OLS or 2SLS), in case we increase, e.g., the number of observations n or the number of parameters k , modify the DGPs and/or hyper-parameter tuning settings.

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 models otherwise. We use random forest models as they are standard choices in the examples from [1] and [2], can fit nonlinear functional forms and deal with collinearity of input variables. As hyper-parameters, we use in the first replication a five-fold grid search across 100 and 400 trees, a maximum of 10 or 20 features, a maximum depth of 5 or arbitrary and a minimum sample of 1 or 4 per leave.

We use also ML algorithms other than random forest models as robustness checks, namely Gradient Boosting and Lasso models. We find that the outcomes of these robustness checks overall confirm the following results. See appendix A for details.

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.

$$Y = c + D\theta + X\beta + U,$$

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

$$\begin{aligned} Y &= c_1 + D\theta + X\beta_{sls} + U, \\ W &= c_2 + X\gamma + Z\delta + \zeta + V, \end{aligned}$$

where W is a $(n \times (k + 1))$ matrix that consists of the matrix X with an additional column which is D . c_1 and c_2 are constants. (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 in alignment with the PLR model

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

3.1 Scenario 1

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 $m = 1, \dots, k$ and $j = 1, \dots, k$. 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 causal 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 causal effect.

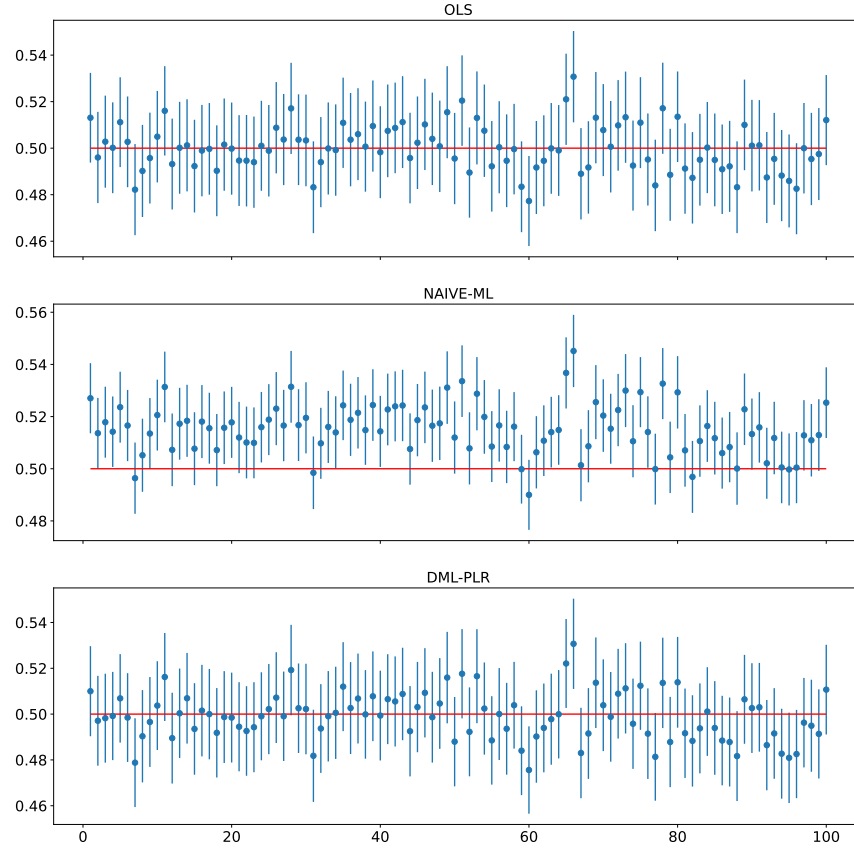


Figure 1: True and estimated causal effects per replication for Scenario 1. The error bars represent the upper and lower bound of the 95% confidence interval of the estimate. The red line represents the true value of the causal effect.

Table A shows the root mean squared error (RMSE), the mean absolute error (MAE) and the bias of estimated causal 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	0.0098	0.0077	0.0000
NAIVE-ML	0.0180	0.0156	0.0152
DML-PLR	0.0104	0.0082	-0.0010
Best	OLS	OLS	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.

We find in Scenario 1 that the DML-PLR and the OLS model tend to recover the causal effect overall similarly in terms of how accurate the point estimates are, of how wide the confidence intervals are and in terms of the (absolute) bias. The OLS model is however slightly better regarding RMSE, MAE and bias than the DML-PLR model. The NAIVE-ML model performs worse in the RMSE, and MAE and has a larger bias compared to DML-PLR and OLS.

3.2 Scenario 2

3.2.1 DGP in alignment with the IRM model

The Scenario 2 represents the case of a IRM model and we use the following DGP for $i = 1, \dots, n$. 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 $m = 1, \dots, k$ and $j = 1, \dots, k$. $\mathcal{U}(a, b)$ 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:

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

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

3.2.2 Results

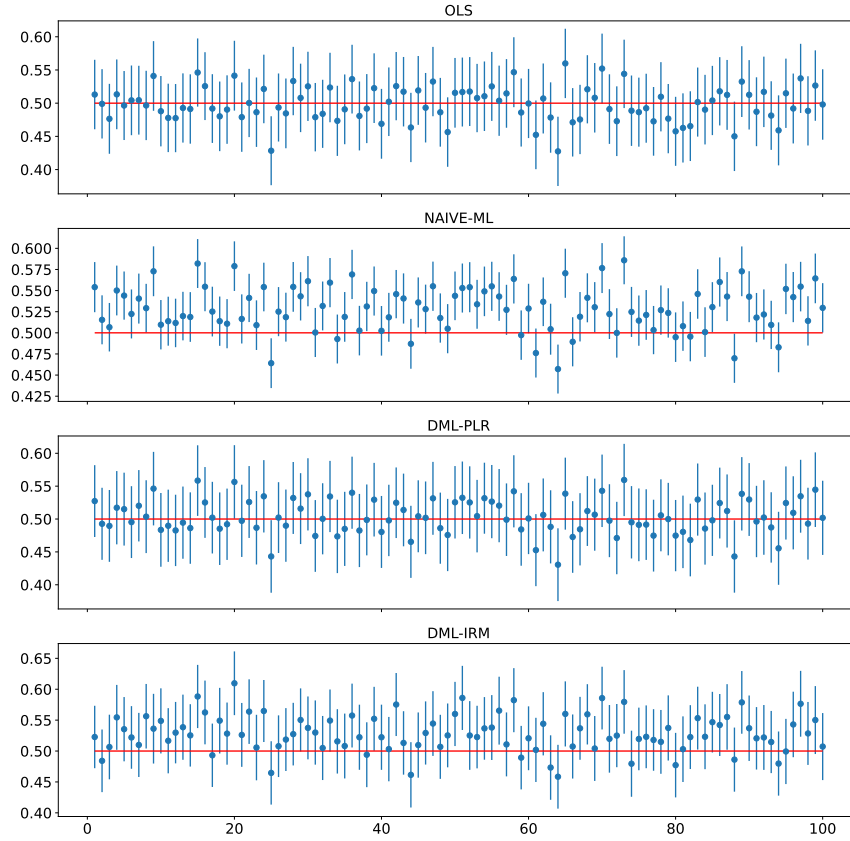


Figure 2: True and estimated causal effects per replication for Scenario 2. The error bars represent the upper and lower bound of the 95% confidence interval of the estimate. The red line represents the true value of the causal effect.

	RMSE	MAE	Bias
OLS	0.0259	0.0209	-0.0011
NAIVE-ML	0.0391	0.0327	0.0288
DML-PLR	0.0263	0.0209	0.0037
DML-IRM	0.0411	0.0338	0.0286
Best	OLS	OLS	OLS

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.

We find in Scenario 2 that the DML-PLR and the OLS model tend to recover the causal effect overall similarly in terms of how wider the confidence intervals, the RMSE, and MAE. The OLS model is however slightly better regarding RMSE, MAE and bias than the DML-PLR model. The NAIVE-ML and the DML-IRM models perform worst compared to the other two models regarding RMSE, MAE and bias. The DML-IRM model is slight worse in terms of RMSE and MAE and slightly better in terms of bias than the NAIVE-ML model.

3.3 Scenario 3

3.3.1 DGP in alignment with the PLR-IV

The Scenario 3 represents the case of a PLR-IV model and we use the following DGP for $i = 1, \dots, n$ [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 $m = 1, \dots, k$ and $j = 1, \dots, k$. I_l is the $l \times l$ identity matrix. $\beta = \gamma$ is a k -vector with entries $\beta_j = \frac{1}{j^2}$ with $j = 1, \dots, k$. δ is a l -vector with entries $\delta_j = \frac{1}{h^2}$, with $h = 1, \dots, l$. 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 . Note also that the DGP is linear in the variables and therefore we expect the linear model 2SLS to perform well.

3.3.2 Results

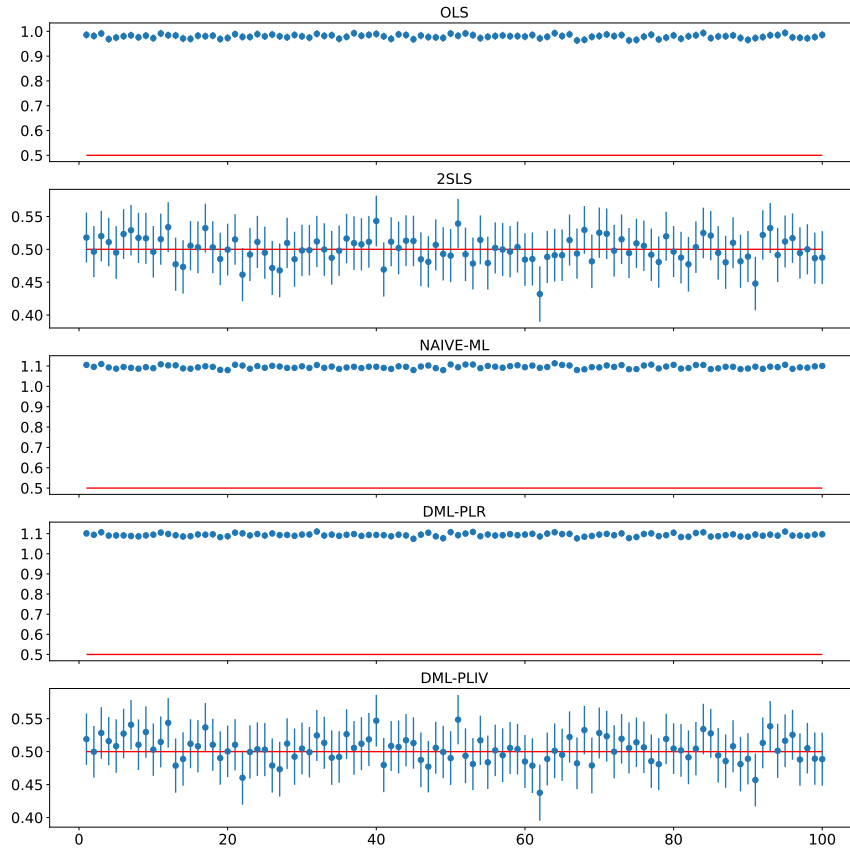


Figure 3: True and estimated causal effects per replication for Scenario 3. The error bars represent the upper and lower bound of the 95% confidence interval of the estimate. The red line represents the true value of the causal effect.

	RMSE	MAE	Bias
OLS	0.4798	0.4797	0.4797
2SLS	0.0189	0.0149	0.0002
NAIVE-ML	0.5952	0.5952	0.5952
DML-PLR	0.5936	0.5935	0.5935
DML-PLIV	0.0199	0.0156	0.0038
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.

We find in Scenario 3 that the 2SLS and the DML-PLIV model tend to recover the causal effect overall similarly in terms of how wider the confidence intervals, the RMSE, MAE and bias. The 2SLS model is however slightly better in all three performance metrics, i.e., RMSE, MAE and bias, than the DML-PLIV model. Note again that the DGP in Scenario 3 is linear in the variables and therefore it makes sense that the 2SLS model performs well and better than the DML-PLIV model. The reason for this outcome is that the 2SLS model assumes already the correct functional form and therefore is subject to less variation in the estimation.

The OLS, NAIVE-ML and DML-PLR models - that do not account for endogeneity of the treatment variable - perform considerably worse regarding RMSE, MAE and bias, whereas OLS is better than the other two.

3.4 Scenario 4

3.4.1 DGP in alignment with the IIV model

The Scenario 4 represents the case of a IIV model and we use the following DGP for $i = 1, \dots, n$. 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 $\text{Bernoulli}(p)$ represents the Bernoulli distribution with parameter p . Σ is a matrix with entries $\Sigma_{kj} = 0.5^{|j-m|}$, with $m = 1, \dots, k$ and $j = 1, \dots, k$. β is a $(k \times 1)$ vector with entries $\beta_j = \frac{1}{j^2}$ for $j = 1, \dots, k$ 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

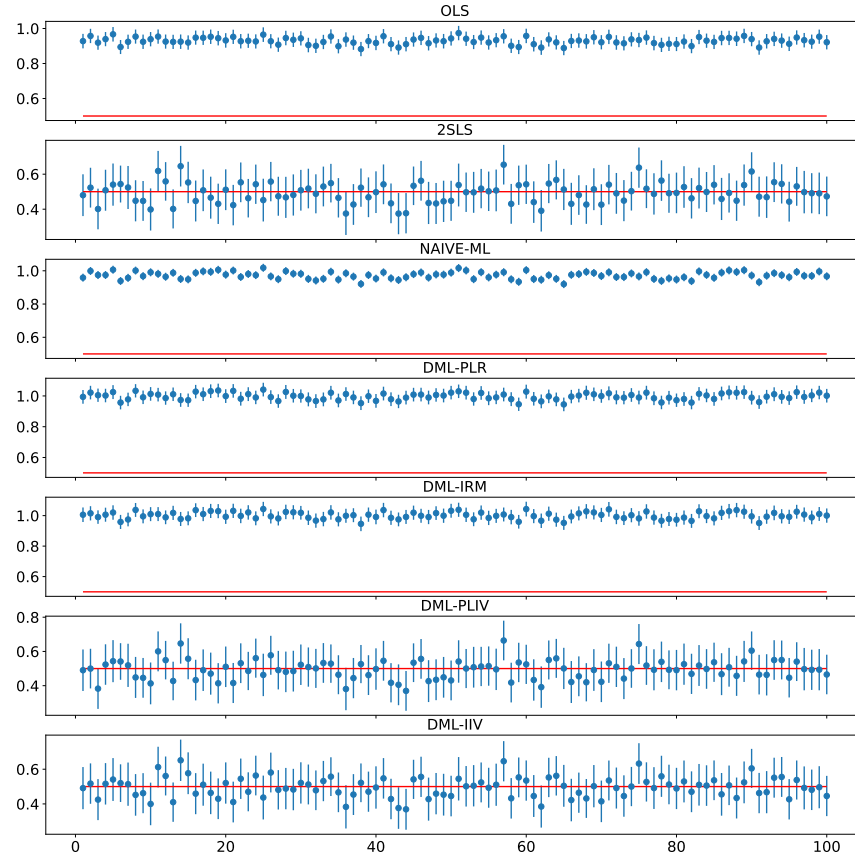


Figure 4: True and estimated causal effects per replication for Scenario 4. The error bars represent the upper and lower bound of the 95% confidence interval of the estimate. The red line represents the true value of the causal effect.

	RMSE	MAE	Bias
OLS	0.4301	0.4297	0.4297
2SLS	0.0574	0.0455	-0.0050
NAIVE-ML	0.4731	0.4726	0.4726
DML-PLR	0.4978	0.4974	0.4974
DML-IRM	0.5014	0.5009	0.5009
DML-PLIV	0.0571	0.0445	-0.0056
DML-IIV	0.0576	0.0458	-0.0041
Best	DML-PLIV	DML-PLIV	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.

We find in Scenario 4 that the DML-IIV, the 2SLS and the DML-PLIV model tend to recover the causal effect overall similarly in terms of how wider the confidence intervals, the RMSE, MAE and bias. The DML-PLIV model is however slightly better in the performance metrics RMSE and MAE and the DML-IIV is better regarding the bias than the other two models.

The OLS, NAIVE-ML, DML-PLR and DML-IRM models - that do not account for endogeneity of the

treatment variable - perform considerably worse across RMSE, MAE and bias, whereas OLS is better than the other three.

4 Discussion and conclusion

We compare the performance of recovering the causal effect parameter in four different model settings and simulation experiments. We considered four widespread DGP settings for DML models in the literature and software implementation. This is either with or without non-linear effects of the causal effect and with and without endogeneity of the associated variable. From the comparison, we consider overall four DML models (i.e., the PLR, IRM, PLR-IV and IIV model), a naive machine learning model and two linear counterparts (i.e., the OLS or 2SLS model). Overall, we find similar performances regarding RMSE, MAE and bias for the appropriate DML model and according linear counterparts (i.e., the OLS or the 2SLS model).

It is at first sight surprising that the linear models (i.e., OLS and 2SLS) perform notable well despite (a) confounded causal effect variables and (b) the non-linearity of the DGP. The outcomes and the reasons for the outcomes are in detail the following.

In the two simulation experiment scenarios without endogeneity of the causal effect variable, we find the naive ML model to perform worst and hence even worse than the simpler OLS model. DML-PLR and OLS model perform best for the linear (i.e., Scenario 1) and non-linear (i.e., Scenario 2) DGPs according to RMSE, MAE and bias. Where overall one of the two, i.e., the DML-PLR and OLS model, performs either slightly better or slightly worse than the other one.

This outcome for Scenario 1 may be due to the fact that a model with a simple linear structure can still recover the linear effect of the treatment variable relatively well even if the nuisance functions are non-linear. Whereas the DML-PLR model can account for the true non-linear structure of the DGP and for the fact that the treatment variable is confounded. However, the DML-PLR model has to trade off its advantageous flexibility with a higher uncertainty of the underlying functional form and hence a decrease in the strength to recover the causal effect.

Similarly in Scenario 2, we presume that the outcome is due to the trade off of the (correct) DML-IRM model between greater flexibility of the estimated functional structure and the decrease in the efficiency to estimate the causal effect with less information in the data. This is, the treatment variable is binary in Scenario 2 and therefore is less informative than, e.g., a ratio scaled variable.

In the two simulation experiment scenarios with endogeneity of the causal effect variable (i.e. Scenarios 3 and 4), we have a linear DGP in one scenario (Scenario 3) and therefore observe the 2SLS model as performing best as it assumes already the correct functional form and does not have to learn it from the data, opposed the DML-PLIV model.

In the other scenario (Scenario 4), with additional non-linear effects of the causal effect variable, we observe a notable better performance of the DML-IIV model compared to the 2SLS and DML-PLIV models. In opposite to Scenario 2, Scenario 4 has despite the more complex model an additional data source, i.e., the instrumental variable, which help to inform the estimation of the non-linear functional form of the model.

The positive result for practitioners and researchers of DML models is that the DML models perform at least similarly well as the simple regression models in the considered simulation experiments (in case we correctly account for the endogeneity of the treatment variable if needed). Therefore, these results suggest that even as we find that the DML models do not always outperform the simple linear regression models, practitioners and researchers may not necessarily face severe disadvantages in recovering the causal effects when using the DML models for their analysis. We found however one minor but noteworthy exception to this conclusion in the Scenario 2, where the DML-PLR model performs on par with the OLS model, and not the DML-IRM model which is in alignment with the IRM model.²

Future research may further analyze the performance differences of DML and simple linear regression models with different settings with more complex DGPs. For instance, if a higher degree on non-linearity is added to the DGPs. Doing so may help to show case the presumed advantage of the DML models over simple linear regression model to recover causal effects.

²As discussed above, this outcome may be due to the fact that the treatment variable is binary in Scenario 2 and therefore is less informative to inform the non-linear machine learning model without additional information from, e.g., instrumental variables.

Appendix

A Results with alternative ML models

We use also ML algorithms other than random forest models as robustness checks, namely Gradient Boosting and Lasso models. We find that the outcomes of these robustness checks overall confirm the results from above in the main text. All scripts for the analysis in this github repository: <https://github.com/g-r-m-n/dml>

A.1 Scenario 1

	RMSE	MAE	Bias
NAIVE-ML Lasso	0.0285	0.0267	0.0267
DML-PLR Lasso	0.0101	0.0081	0.0021
NAIVE-ML XGBoost	0.0133	0.0109	0.0090
DML-PLR XGBoost	0.0101	0.0081	-0.0006

Table A: 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.

A.2 Scenario 2

	RMSE	MAE	Bias
NAIVE-ML Lasso	0.0441	0.0373	0.0332
DML-PLR Lasso	0.0290	0.0220	-0.0024
DML-IRM Lasso	0.0364	0.0278	0.0026
NAIVE-ML XGBoost	0.0316	0.0248	0.0097
DML-PLR XGBoost	0.0304	0.0232	-0.0007
DML-IRM XGBoost	0.0367	0.0283	0.0130

Table B: 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.

A.3 Scenario 3

	RMSE	MAE	Bias
NAIVE-ML Lasso	0.6023	0.6022	0.6022
DML-PLR Lasso	0.5992	0.5992	0.5992
DML-PLIV Lasso	0.0207	0.0166	0.0051
NAIVE-ML XGBoost	0.5953	0.5953	0.5953
DML-PLR XGBoost	0.5941	0.5940	0.5940
DML-PLIV XGBoost	0.0210	0.0167	0.0025

Table C: 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.

A.4 Scenario 4

	RMSE	MAE	Bias
NAIVE-ML Lasso	0.4779	0.4773	0.4773
DML-PLR Lasso	0.5020	0.5014	0.5014
DML-IRM Lasso	0.5080	0.5074	0.5074
DML-PLIV Lasso	0.0630	0.0495	-0.0010
DML-IIV Lasso	0.0617	0.0485	-0.0017
NAIVE-ML XGBoost	0.4711	0.4706	0.4706
DML-PLR XGBoost	0.4979	0.4973	0.4973
DML-IRM XGBoost	0.4944	0.4938	0.4938
DML-PLIV XGBoost	0.0627	0.0496	-0.0026
DML-IIV XGBoost	0.0632	0.0505	-0.0044

Table D: 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.

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.