Is Double Machine Learning always better than Linear Regression to estimate Causal Effects?

Evidence from four simulation experiments

Dr. German Zenetti

8. April 2023

Personal note

Unfortunately, I'm legally obligated to *not* show any details, results and/or code of my work for my previous employer

Therefore, I will present a recent private project instead.

Presentation Outline

1. Topic

Research question

Is Double Machine Learning (DML) always better than Simple Linear Regression to estimate Causal Effects?

Evidence from four simulation experiments

What is DML?

Partially linear regression (**PLR**) model (Chernozhukov et al. (2015)):

$$
\begin{aligned}
Y &= D\theta_0 + g_0(X) + U, \quad E[U|X,D] = 0 &\quad (1)\\
D &= m_0(X) + V, \quad E[V|X] = 0 &\quad (2)
\end{aligned}
$$

Y: $(n \times 1)$ outcome variable.
D: $(n \times 1)$ treatment.
X: $(n \times k)$ matrix of k confounding or control variables.
g_0, m_0: nuisance functions mapping X to $\mathbb{R}$.
U, V: $(n \times 1)$ error term vectors.
θ_0: scalar parameter of central interest.
$E[V|X]$: expected value of variable V given variables X.

Basic idea of approach

Naive approach:
Iteratively estimate g_0 and θ until convergence using Equation (1) only. $\rightarrow$ Bias due to regularization and overfitting (as D depends on X).

Idea: Partialling-out:

1. Estimate $D = \hat{m}_0(X) + \hat{V}$ (treatment model).
2. Estimate $Y = D\hat{\theta}_0 + \hat{g}_0(X) + \hat{U}$ (outcome model) using the naive approach.
3. Regress $Y - \hat{g}_0(X)$ on $\hat{V}$ using ordinary least squares (OLS) to estimate θ.
4. Estimated θ is free of regularization bias.

Use this idea of orthogonalization to formulate Neyman Orthogonality conditions together with sample splitting.

2. Simulation Experiments

Overview

Four simulation experiment scenarios:

- with linear or non-linear effects
- with or without endogeneity of causal effect variable

Based on widespread data generating processes for DML models in literature and software.

In each simulation experiment:
$n = 10.000$ observations of data,
$k = 20$ variables,
$\theta = 0.5$,
Number of replication $= 100$.
Tuning: use first replication for 5-fold grid search

Similar outcomes: if, e.g., n or k, modify the parameters in DGPs or hyper-parameter tuning settings.

Simulation Experiment PLR: Setting

PLR model data generating process from Chernozhukov et al. (2018):

- with linear effects and
- without endogeneity of causal effect variable

Simulation Experiment PLR: Setting I

PLR model DGP fromChernozhukov et al. (2018):

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

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

Simulation Experiment PLR: Setting II

$\sigma_{mj} = 0.7^{|j-m|}$, 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.

Simulation Experiment PLR: Results

Tabelle: Root mean squared error (RMSE), mean absolute error (MAE) and bias of estimated treatment effect and the true value across the replications.

	RMSE	MAE	Bias
OLS	0.0098	0.0077	0.0000
NAIVE-ML-RF 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
NAIVE-ML-RF	0.0180	0.0156	0.0152
DML-PLR-RF	0.0104	0.0082	-0.0010
Best	OLS	OLS	OLS

Simulation Experiment PLR-IV: Setting

PLR-IV (partial linear regression instrumental variables) model data generating process from Belloni et al. (2017):

- with linear effects and
- with endogeneity of causal effect variable

Simulation Experiment PLR-IV: Results

Tabelle: RMSE, MAE and bias of estimated treatment effect and the true value across the replications.

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

Simulation Experiment IRM: Setting

IRM (interactive regression model) model data generating process from Chernozhukov et al. (2015):

- with non-linear effects and
- without endogeneity of causal effect variable

Simulation Experiment IRM: Results

Tabelle: RMSE, MAE and bias of estimated treatment effect and the true value across the replications.

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

Simulation Experiment IIV: Setting

IIV (interactive regression model using instrumental variables) model data generating process from Chernozhukov et al. (2015):

- with non-linear effects and
- with endogeneity of causal effect variable

Simulation Experiment IIV: Results

Tabelle: RMSE, MAE and bias of estimated treatment effect and the true value across the replications.

	RMSE	MAE	Bias
OLS	0.4301	0.4297	0.4297
2SLS	0.0574	0.0455	-0.0050
DML-PLIV Lasso	0.0630	0.0495	-0.0010
DML-IIV Lasso	0.0617	0.0485	-0.0017
DML-PLIV XGBoost	0.0627	0.0496	-0.0026
DML-IIV XGBoost	0.0632	0.0505	-0.0044
NAIVE-ML-RF	0.4731	0.4726	0.4726
DML-PLR-RF	0.4978	0.4974	0.4974
DML-IRM-RF	0.5014	0.5009	0.5009
DML-PLIV-RF	0.0571	0.0445	-0.0056
DML-IIV-RF	0.0576	0.0458	-0.0041
Best	DML-PLIV	DML-PLIV	DML-IIV

3. Conclusion

Conclusion

- Compared 4 DML models (PLR, IRM, PLR-IV and IIV), a naive ML model and two linear counterparts (OLS or 2SLS)
- Similar performances (RMSE, MAE and bias) for appropriate DML model and linear model
- despite (a) confounded causal effect variables and (b) non-linearity of the DGP
- DML model trades off flexibility with higher uncertainty of underlying functional form and decrease in strength to recover causal effect
- Naive ML typically performs best for predicting Y
- DML-models outperform OLS and 2SLS models when we add additional non-linearity of the DGP

4. References

References I

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

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., Hansen, C., and Spindler, M. (2015). Post-selection and post-regularization inference in linear models with many controls and instruments. *American Economic Review*, 105(5):486–90.

Farbmacher, H., Guber, R., and Klaaßen, S. (2020). Instrument validity tests with causal forests. *MEA Discussion Paper*, 13.

5. Appendix

OLS model

Linear regression model, denoted as OLS of the following form:

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

where c is a constant.

2SLS model

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.

Naive ML model

Naive ML model, denoted as naive-ML of the following form:

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

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

Simulation Experiment IRM: Setting I

The DGP of the IRM model based on Belloni et al. (2017).

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

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

Simulation Experiment IRM: Setting II

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.

Simulation Experiment PLR-IV: Setting I

The DGP of PLR-IV model based on Chernozhukov et al. (2015).

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

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

Simulation Experiment PLR-IV: Setting II

k-vector with entries $\beta_j = \frac{1}{j^2}$ with $j = 1, \ldots, k$. δ is a l-vector with entries $\delta_j = \frac{1}{h^2}$, with $h = 1, \ldots, l$. 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.

Simulation Experiment IIV: Setting I

The DGP of the IIV model based on Farbmacher et al. (2020).

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

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

Simulation Experiment IIV: Setting II

$m = 1, .., k$ 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.