DOTE 6635: Artificial Intelligence for Business Research

Double Machine Learning

Renyu (Philip) Zhang

1

Machine Learning for Causal Inference

- Using machine learning for causal inference is generally very challenging.
  - Cross-validation cannot be directly applied to hyper-parameter tuning for causal inference models (Athey and Imbens, 2016).
  - Good performance in predicting propensity scores or outcomes cannot be directly translated into good causal performance (Belloni et al., 2014).
  - Regularization in ML will introduce additional biases in causal inference (Belloni et al., 2016).
- What are the most critical issues in causal inference at large?
  - Confoundedness, non-overlapping, balance, etc.
  - AI/ML cannot magically solve these fundamental problems of causal inference.
- Double machine learning (DML) provides a framework to empower causal inference with ML.
  - Compared with other fields in business research, this is a very fast-evolving field of study.

Recursive partitioning for heterogeneous **causal** effects
S Athey, G Imbens - Proceedings of the National Academy of Sciences, 2016 - pnas.org
… We refer to the estimators developed in this section as "**causal tree**" (CT) estimators. … for constructing **trees** for **causal** effects that allow us to do valid inference for the **causal** effects in …
☆ Save 99 Cite Cited by 2213 Related articles All 17 versions »

Inference on treatment effects after selection among high-dimensional controls
A Belloni, V Chernozhukov… - Review of Economic …, **2014** - academic.oup.com
We propose robust methods for inference about the effect of a treatment variable on a scalar outcome in the presence of very many regressors in a model with possibly non-Gaussian …
☆ Save 99 Cite Cited by 2040 Related articles All 32 versions »

Post-selection inference for generalized linear models with many controls
A Belloni, V Chernozhukov, Y Wei - Journal of Business & …, **2016** - Taylor & Francis
This article considers generalized linear models in the presence of many controls. We lay out a general methodology to estimate an effect of interest based on the construction of an …
☆ Save 99 Cite Cited by 233 Related articles All 11 versions Web of Science: 87 »

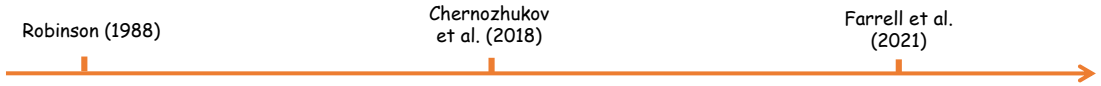
2

2

Today's Focus

Root-N-consistent semiparametric regression
PM **Robinson** - Econometrica: Journal of the Econometric Society, **1988** - JSTOR
One type of semiparametric regression on an $\scr{R}^{p}\times \scr{R}^{q}\text{-valued}$ random variable (X, Z) is β'X + θ(Z), where β and θ(Z) are an unknown slope coefficient …
☆ Save 99 Cite Cited by 3539 Related articles All 11 versions Web of Science: 1350 »

Deep learning for **individual heterogeneity**: An **automatic inference framework**
MH Farrell, T Liang, S Misra - arXiv preprint arXiv:2010.14694, 2020 - arxiv.org
… and **inference** using **machine learning** to enrich economic models. Our **framework** takes a … functions, to capture the rich **heterogeneity** based on potentially high dimensional or complex …
☆ Save 99 Cite Cited by 65 Related articles All 13 versions »



Partial Linear Models — Double Machine Learning — DML in Action

Double/debiased machine learning for treatment and structural parameters
V **Chernozhukov**, D Chetverikov, M Demirer, E Duflo… - **2018** - academic.oup.com
We revisit the classic semi-parametric problem of inference on a low-dimensional parameter θ 0 in the presence of high-dimensional nuisance parameters η 0 . We depart from the …
☆ Save 99 Cite Cited by 3486 Related articles All 29 versions Web of Science: 1112 »

Applied DML: A Historical Perspective

Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf

3

3

High-Level Takeaways of the DML Literature

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf
DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

- Provides a general framework, by leveraging Neyman Orthogonality, for estimating treatment effects using ML methods.
  - Causal inference usually requires estimating the. expected outcomes (and propensity scores) conditioned on covariates or confounders
  - Standard econometrics methods make strong functional form assumptions (e.g., linear models), which require strong substantive justifications; if mis-specified, causal estimates will be significantly biased.
  - DML framework automatically learns the form of conditional expectation functions from data.
- DML framework requires:
  - Some regularity conditions (recall the assumptions for AIPW)
  - ML estimators to converge, in RMSE/L2-norm at a rate of $o(n^{-1/4})$, slower than $o(n^{-1/2})$, the rate of most parametric models according to the delta method.
- DML framework outputs:
  - Root-n consistent estimators for treatment effects: Convergence to the ground-truth in probability at a rate $O(n^{-1/2})$, a property natural and common in a parametric world.
  - In frequentist perspective, root-n consistency typically means asymptotically normal, which means you can construct valid confidence intervals and do inference on your estimators.

4

4

Agenda

- Partial Linear Models
- General Double Machine Learning Framework
- Double Machine Learning in Action: Practical Recipe and Pitfalls

5

5

Let's First Look at a Simple Model

Causal Inference: A Statistical Learning Approach https://web.stanford.edu/~swager/causal_inf_book.pdf
Applied Causal Inference Powered by ML and AI: https://chapters.causalml-book.org/CausalML_book_2022.pdf
DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/torugf4016/files/bstewart/files/chern.handout.pdf

- We use the partial linear model (PLM) to illustrate the DML framework:
  - Why ML is important and useful;
  - How the statistical theory works.

Partially Linear Model Set-up

- Y: Outcome
- D: Treatment
- X: Measured confounders
- U and V are our error terms
- We assume zero conditional mean:

$$Y = D\theta_0 + g_0(X) + U$$

$$D = m_0(X) + V$$

$$\mathrm{E}[U \mid X, D] = 0 \qquad \mathrm{E}[V \mid X] = 0$$

6

6

Partial Linear Model

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

Partially Linear Model Set-up

$$Y = D\theta_0 + g_0(X) + U$$
$$D = m_0(X) + V$$

- Y: Outcome
- D: Treatment
- X: Measured confounders
- U and V are our error terms
- We assume zero conditional mean:

$$\mathrm{E}[U \mid X, D] = 0 \qquad \mathrm{E}[V \mid X] = 0$$

- θ_0 is the parameter of interest.
- $g_0(.)$, the and $m_0(.)$ can take any arbitrary functional forms.
- X can be very high-dimensional, potentially dim(X) >> dim (D).
- Here, linearly additive separability is simply for illustration purposes.

7

7

Partial Linear Model

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

$$Y = D\theta_0 + g_0(X) + U$$
$$D = m_0(X) + V$$
$$\mathrm{E}[U \mid X, D] = 0 \qquad \mathrm{E}[V \mid X] = 0$$

- PLM: High-dimensional but observable confounders.
- If $g_0(.)$ and $m_0(.)$ are known, we can estimate θ_0 with a root-n consistent estimator.
- But $g_0(.)$ and $m_0(.)$ are unknown in practice, and this is where ML plays a crucial role.
- ML relaxes the linearity and additivity of $g_0(.)$ and $m_0(.)$, but ML alone is not sufficient (just like IPW is not sufficient).
- We can also think of PLM from the perspective of propensity scores:
  - $m_0(X)$ can be thought of as the propensity score.
  - If we can have an estimator $\hat{m}_0(.)$ for $m_0(.)$, can we construct an AIPW-type estimator for θ_0?

8

8

Partial Linear Model: Naïve Approach

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

- Two sources of biases: Regularization bias and overfitting bias.
- Regularization bias: ML model cannot predict the functions sufficiently fast.
  - Addressed by Neyman Orthogonality, or orthogonal moment conditions.
- Overfitting bias: The ML estimator may overfit the data with which it is trained.
  - Addressed by cross-fitting, i.e., splitting the data for training and inference.

$$Y = D\theta_0 + g_0(X) + U$$
$$D = m_0(X) + V$$
$$E[U \mid X, D] = 0 \qquad E[V \mid X] = 0$$

Regularization bias. A naive approach to estimation of θ_0 using ML methods would be, for example, to construct a sophisticated ML estimator $D\hat{\theta}_0 + \hat{g}_0(X)$ for learning the regression function $D\theta_0 + g_0(X)$.[2] Suppose, for the sake of clarity, that we randomly split the sample into two parts: a main part of size n, with observation numbers indexed by $i \in I$, and an auxiliary part of size $N - n$, with observations indexed by $i \in I^c$. For simplicity, we take $n = N/2$ for the moment and we turn to more general cases that cover unequal split-sizes, using more than one split, and achieving the same efficiency as if the full sample were used for estimating θ_0 in the formal development in Section 3. Suppose $\hat{g}_0$ is obtained using the auxiliary sample and that, given this $\hat{g}_0$, the final estimate of θ_0 is obtained using the main sample:

$$\hat{\theta}_0 = \left(\frac{1}{n}\sum_{i\in I} D_i^2\right)^{-1} \frac{1}{n}\sum_{i\in I} D_i(Y_i - \hat{g}_0(X_i)).$$

(1.3) ☹ $\Rightarrow |\sqrt{n}(\hat{\theta}_0 - \theta_0)| \xrightarrow{p} \infty$

9

9

Regularization Bias of Naïve Approach

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

$$Y = D\theta_0 + g_0(X) + U$$
$$D = m_0(X) + V$$
$$E[U \mid X, D] = 0 \qquad E[V \mid X] = 0$$

Regularization bias. A naive approach to estimation of θ_0 using ML methods would be, for example, to construct a sophisticated ML estimator $D\hat{\theta}_0 + \hat{g}_0(X)$ for learning the regression function $D\theta_0 + g_0(X)$.[2] Suppose, for the sake of clarity, that we randomly split the sample into two parts: a main part of size n, with observation numbers indexed by $i \in I$, and an auxiliary part of size $N - n$, with observations indexed by $i \in I^c$. For simplicity, we take $n = N/2$ for the moment and we turn to more general cases that cover unequal split-sizes, using more than one split, and achieving the same efficiency as if the full sample were used for estimating θ_0 in the formal development in Section 3. Suppose $\hat{g}_0$ is obtained using the auxiliary sample and that, given this $\hat{g}_0$, the final estimate of θ_0 is obtained using the main sample:

$$\hat{\theta}_0 = \left(\frac{1}{n}\sum_{i\in I} D_i^2\right)^{-1} \frac{1}{n}\sum_{i\in I} D_i(Y_i - \hat{g}_0(X_i)).$$

(1.3) ☹ $\Rightarrow |\sqrt{n}(\hat{\theta}_0 - \theta_0)| \xrightarrow{p} \infty$

$$\sqrt{n}(\hat{\theta}_0 - \theta_0) = \underbrace{\left(\frac{1}{n}\sum_{i\in I} D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i\in I} D_i U_i}_{:=a} + \underbrace{\left(\frac{1}{n}\sum_{i\in I} D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i\in I} D_i(g_0(X_i) - \hat{g}_0(X_i))}_{:=b}$$

- Oracle estimator with known $g_0(.)$
- Asymptotically normal by linear regression or CLT.

$$b = (E[D_i^2])^{-1} \frac{1}{\sqrt{n}}\sum_{i\in I} m_0(X_i)(g_0(X_i) - \hat{g}_0(X_i)) + o_P(1)$$

If the RMSE of $\hat{g}_0(.)$ is of order $n^{-\varphi_g}$, where $\varphi_g < 1/2$, b will be of order $\sqrt{n}n^{-\varphi_g} \to \infty$.

10

10

Overcoming Regularization Bias

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

$$\hat{\theta}_0 = \left(\frac{1}{n}\sum_{i\in I} D_i^2\right)^{-1} \frac{1}{n}\sum_{i\in I} D_i(Y_i - \hat{g}_0(X_i))$$

$$Y = D\theta_0 + g_0(X) + U$$

$$D = m_0(X) + V$$

$$\mathrm{E}[U \mid X, D] = 0 \qquad \mathrm{E}[V \mid X] = 0$$

$$\sqrt{n}(\hat{\theta}_0 - \theta_0) = \underbrace{\left(\frac{1}{n}\sum_{i\in I} D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i\in I} D_i U_i}_{:=a} + \underbrace{\left(\frac{1}{n}\sum_{i\in I} D_i^2\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i\in I} D_i(g_0(X_i) - \hat{g}_0(X_i))}_{:=b}$$

$$b = (E[D_i^2])^{-1} \frac{1}{\sqrt{n}}\sum_{i\in I} m_0(X_i)(g_0(X_i) - \hat{g}_0(X_i)) + o_P(1)$$

- Orthogonality will help get ride of the bias in b.
  - Not a new idea, which was proposed in Robinson (1988) and Frisch-Waugh-Lovell (FWL) Theorem (1930).
  - Also proposed in the authors' prior work, Belloni et al. (2014).
- The basic idea is to train two ML models and double select out the biases, called double selection in Belloni et al. (2014).

Partial time regressions as compared with **individual trends**
R Frisch, FV Waugh - Econometrica: Journal of the Econometric Society, 1933 - JSTOR
… allowed to influence the **regression** coefficient between … **partial time regression** method instead of the **individual trend** method. This is false. The possibility of determining the long **time** …
☆ Save 99 Cite Cited by 766 Related articles All 4 versions »»

Root-N-consistent semiparametric regression
PM **Robinson** - Econometrica: Journal of the Econometric Society, **1988** - JSTOR
One type of semiparametric regression on an $\scr{R}^{p}\times \scr{R}^{q}\text{-valued}$ random variable (X, Z) is β'X + θ(Z), where β and θ(Z) are an unknown slope coefficient …
☆ Save 99 Cite Cited by 3539 Related articles All 11 versions Web of Science: 1350 »»

Inference on treatment effects after selection among high-dimensional controls
A **Belloni**, V **Chernozhukov**… - Review of Economic …, **2014** - academic.oup.com
We propose robust methods for inference about the effect of a treatment variable on a scalar outcome in the presence of very many regressors in a model with possibly non-Gaussian …
☆ Save 99 Cite Cited by 2040 Related articles All 32 versions »»

11

11

Frisch-Waugh-Lovell (FWL) Theorem

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

- Let's say we want to estimate the following model using OLS:
  - $Y = \beta_0 + \beta_1 D + \beta_2 X + U$
- The Frisch–Waugh–Lovell Theorem shows us that we can recover the OLS estimate of β_1 using a residuals-on-residuals OLS regression:
  1. Regress D on X using OLS
     - Let $\widehat{D}$ be the predicted values of D and let the residuals $\widehat{V} = D - \widehat{D}$
  2. Regress Y on X using OLS
     - Let $\widehat{Y}$ be the predicted values of Y and let the residuals $\widehat{W} = Y - \widehat{Y}$
  3. Regress $\widehat{W}$ on $\widehat{V}$ using OLS

Residual-on-residual regression.

- The estimated coefficient on $\widehat{V}$ will be the same as the estimated coefficient $\hat{\beta}_1$ from regressing Y on D and X using OLS!

12

12

Robinson (1988)

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

- The **Frisch–Waugh–Lovell** procedure:
  1. **Linear** regression of D on X
  2. **Linear** regression of Y on X
  3. **Linear** regression of the residuals from 2 on the residuals from 1
- Robinson's innovation: let's replace the **linear** regressions from 1 and 2 with some **non-parametric** regression
- **Robinson's** procedure:
  1. **Kernel** regression of D on X
  2. **Kernel** regression of Y on X
  3. **Linear** regression of the residuals from 2 on the residuals from 1

Neural nets, tree-based models, kernel regressions, etc.

13

13

Double Machine Learning (2018)

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

- The idea proposed in Chernozhukov et al. (2018) is similar to Robinson (1988).

- DML using residuals-on-residuals regression:
  1. Estimate $D = \hat{m}_0(X) + \hat{V}$
  2. Estimate $Y = \hat{\ell}_0(X) + \hat{U}$
     - Note the absence of D and the switch from $g_0(\cdot)$ to $\ell_0(\cdot)$, which is essentially $E[Y \mid X]$
  3. Regress $\hat{U}$ on $\hat{V}$ using OLS for an estimate $\check{\theta}_0$

DML Estimator

$$\check{\theta}_0 = \left(\frac{1}{n}\sum_{i\in I}\hat{V}_i\hat{V}_i\right)^{-1}\frac{1}{n}\sum_{i\in I}\hat{V}_i(Y_i - \hat{\ell}_0(X_i)), \quad \ell_0(X) = E[Y|X]$$

- **Robinson's** procedure:
  1. Predict D with X using **kernel regression**
  2. Predict Y with X using **kernel regression**
  3. **Linear** regression of the residuals from 2 on the residuals from 1
- **DML residuals-on-residuals** procedure:
  1. Predict D with X using **any $n^{\frac{1}{4}}$-consistent ML model**
  2. Predict Y with X using **any $n^{\frac{1}{4}}$-consistent ML model**
  3. **Linear** regression of the residuals from 2 on the residuals from 1

Assumption: RMSE of ML models are of order $o(n^{-1/4})$.

14

14

DML: Debias using Orthogonalization

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

- Orthogonalization: Partialling out the effect of X from D.

Overcoming Regularization Biases using Orthogonalization. Now consider a second construction that employs an "orthogonalized" formulation obtained by directly partialling out the effect of X from D to obtain the orthogonalized regressor $V = D - m_0(X)$. Specifically, we obtain $\widehat{V} = D - \widehat{m}_0(X)$, where $\widehat{m}_0$ is an ML estimator of m_0 obtained using the auxiliary sample of observations. We are now solving an auxiliary prediction problem to estimate the conditional mean of D given X, so we are doing "double prediction" or "double machine learning".

After partialling the effect of X out from D and obtaining a preliminary estimate of g_0 from the auxiliary sample as before, we may formulate the following "debiased" machine learning estimator for θ_0 using the main sample of observations:[3]

$$\check{\theta}_0 = \Big(\frac{1}{n}\sum_{i\in I}\widehat{V}_i D_i\Big)^{-1}\frac{1}{n}\sum_{i\in I}\widehat{V}_i(Y_i - \hat{g}_0(X_i)). \tag{1.5}$$

$$\sqrt{n}(\check{\theta}_0 - \theta_0) = a^* + b^* + c^*$$

$$a^* = (E[V^2])^{-1}\frac{1}{\sqrt{n}}\sum_{i\in I} V_i U_i \rightsquigarrow N(0, \Sigma)$$

The remaining term c* converges to 0 if we split samples or cross-fit.

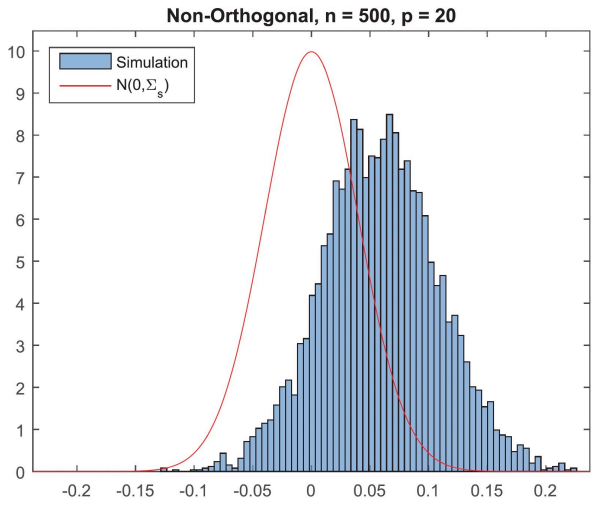
$$b^* = (E[V^2])^{-1}\frac{1}{\sqrt{n}}\sum_{i\in I}(\hat{m}_0(X_i) - m_0(X_i))(\hat{g}_0(X_i) - g_0(X_i)) \leq \sqrt{n}n^{-(\varphi_m+\varphi_g)} \leq o(1)$$

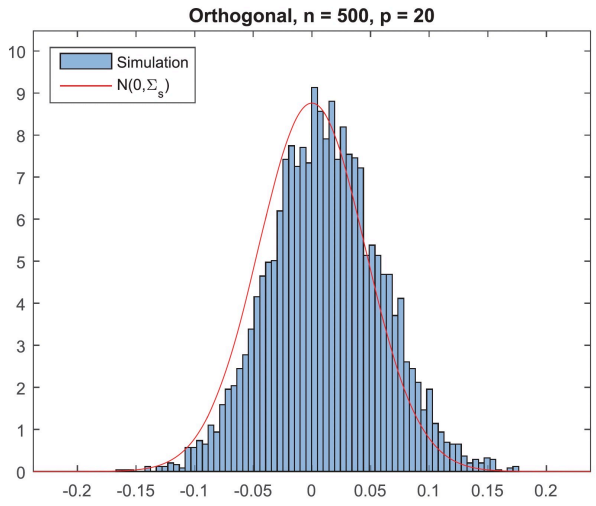
15

15

DML vs. Naïve Plug-in

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf





16

16

Orthogonalization as Instrument

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

- Another way of thinking about orthogonalization is that you create an instrumental variable V on D that is uncorrelated with $Y - g_0(X)$ but correlated with D.

$$Y = D\theta_0 + g_0(X) + U$$
$$D = m_0(X) + V$$
$$E[U \mid X, D] = 0 \qquad E[V \mid X] = 0$$

- DML estimator:

$$\check{\theta}_0 = \left(\frac{1}{n}\sum_{i\in I} \hat{V}_i D_i\right)^{-1} \frac{1}{n}\sum_{i\in I} \hat{V}(Y_i - \hat{g}_0(X_i))$$

- Typical IV:

$$\hat{\beta}_{IV} = (Z'D)^{-1}Z'y$$

17

17

Overcoming Overfitting Bias via Sample Splitting

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

- Recall the overfitting bias.

$$Y = D\theta_0 + g_0(X) + U$$
$$D = m_0(X) + V$$
$$E[U \mid X, D] = 0 \qquad E[V \mid X] = 0$$

- DML estimator: $\check{\theta}_0 = \left(\frac{1}{n}\sum_{i\in I} \hat{V}_i D_i\right)^{-1} \frac{1}{n}\sum_{i\in I} \hat{V}(Y_i - \hat{g}_0(X_i))$

$$\sqrt{n}(\check{\theta}_0 - \theta_0) = a^* + b^* + c^*$$

- We have shown that a* is asymptotically normal, and b* converges to 0 under common convergence rate assumptions for ML models.
- The third part c* contains the following term that only vanishes in probability if we use split samples to estimate $g_0(.)$:

$$\frac{1}{\sqrt{n}}\sum_{i\in I} V_i(\hat{g}_0(X_i) - g_0(X_i)) \qquad (1.6)$$

- With split samples, $V_i(\hat{g}_0(X_i) - g_0(X_i))$ has zero mean and $\frac{1}{n}\sum_{i\in I}(\hat{g}_0(X_i) - g_0(X_i))^2 \xrightarrow{p} 0$. (1.6) will vanish by the Chebyshev's Inequality.

18

18

Split Sample vs. Full Sample

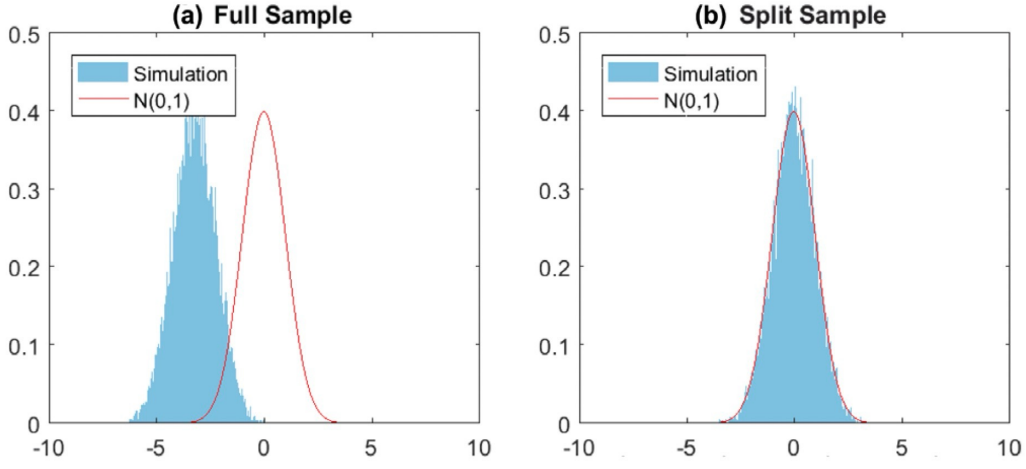
DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

ML Model:

$$\hat{g}_0(X_i) = g_0(X_i) + (Y_i - g_0(X_i))/N^{1/2-\epsilon}$$

Full Sample:

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} V_i(\hat{g}_0(X_i) - g_0(X_i)) \propto N^{\epsilon} \to \infty$$



19

19

Cross-Fitting to Improve Sample Efficiency

DML Package: https://docs.doubleml.org/stable/index.html#
Slides on DML: https://bstewart.scholar.princeton.edu/sites/g/files/toruqf4016/files/bstewart/files/chern.handout.pdf

- Sample splitting reduces bias at the cost of compromised efficiency.
- Like cross-validation, we can use cross-fitting to improve sample efficiency:

$$Y = D\theta_0 + g_0(X) + U$$
$$D = m_0(X) + V$$
$$\mathrm{E}[U \mid X, D] = 0 \qquad \mathrm{E}[V \mid X] = 0$$

1. Randomly partition your data into two subsets
2. Fit two ML models $\hat{g}_{0,1}$ and $\hat{m}_{0,1}$ in the first subset
3. Estimate $\check{\theta}_{0,1}$ in the second subset using the $\hat{g}_{0,1}$ and $\hat{m}_{0,1}$ functions we fit in the first subset
4. Fit two ML models $\hat{g}_{0,2}$ and $\hat{m}_{0,2}$ in the second subset
5. Estimate $\check{\theta}_{0,2}$ in the first subset using the $\hat{g}_{0,2}$ and $\hat{m}_{0,2}$ functions we fit in the second subset
6. Average our two estimates $\check{\theta}_{0,1}$ and $\check{\theta}_{0,2}$ for our final estimate $\check{\theta}_0$

Can be easily generalized to k-fold.

- How about the standard error and confidence interval?
- For PLM, we can just use the SE of stratified estimators; for general DML, we will talk about it later.

20

20

DML in Action

What, Why, and How: An Empiricist's Guide to Double/Debiased Machine Learning

Research Commentary

Bowen Shi
School of Economics and Management, Tsinghua University, sbw22@mails.tsinghua.edu.cn
Xiaojie Mao
School of Economics and Management, Tsinghua University, maoxj@sem.tsinghua.edu.cn
Mochen Yang
Carlson School of Management, University of Minnesota, yang3653@umn.edu
Bo Li
School of Economics and Management, Tsinghua University, libo@sem.tsinghua.edu.cn

This research commentary introduces Double/Debiased Machine Learning (DML), a novel methodological framework, to the Information Systems (IS) research community, demonstrating its power to address the challenges of empirical model specifications. DML combines the flexibility of modern machine learning (ML) techniques with the rigor of semiparametric statistical theory, enabling effective modeling of complex functions alongside valid statistical inference. The paper provides an accessible and comprehensive overview of DML's key elements—Neyman Orthogonality, cross-fitting, and high-quality ML estimation—and their roles in achieving methodological flexibility and rigor. The versatility of DML is illustrated through applications in several empirical settings common in IS research, including standard linear regression with control covariates, instrumental variable regressions, difference-in-differences, and scenarios with ML-generated covariates. Comparative simulations and real data analyses show that DML outperforms traditional parametric and semiparametric methods, and also illustrate the importance of DML's key elements. Finally, we highlight potential misconceptions and pitfalls in applying DML and offer practical advice for empirical researchers. Given the increasing complexity of data and research questions in the IS field, DML offers a timely and powerful tool for empirical researchers. By promoting a deeper understanding and appropriate use of DML, this commentary aims to empower empirical research in IS.

Key words: double/debiased machine learning, model misspecification, statistical inference, semiparametric model, empirical methods

An (IS) Empiricist's Guide to DML:
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4677153

Promotion of DML in the IS community.

Demonstrations of DML in linear models: OLS, IV, DiD, etc.

Discussions of potential misconceptions and pitfalls in applying DML.

What, Why, and How: An Empiricist's Guide to Double/Debiased Machine Learning
B Shi, X Mao, M Yang, B Li - Debiased Machine Learning …, 2023 - papers.ssrn.com
… introduces Double/Debiased Machine Learning (DML), a … the flexibility of machine learning techniques with the rigor of … , and scenarios with machine learning-generated covariates. …
☆ 保存 99 引用 被引用次数：2 相关文章 »

21

21

Evaluate Persuasive Power of Reputation

informs
https://pubsonline.informs.org/journal/mnsc
MANAGEMENT SCIENCE
Vol. 70, No. 3, March 2024, pp. 1613–1634
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Influence via Ethos: On the Persuasive Power of Reputation in Deliberation Online

Emaad Manzoor,[a,*] George H. Chen,[b] Dokyun Lee,[c] Michael D. Smith[b]
[a] Cornell University, Ithaca, New York 14850; [b] Carnegie Mellon University, Pittsburgh, Pennsylvania 15213; [c] Boston University, Boston, Massachusetts 02215
*Corresponding author
Contact: emaadmanzoor@cornell.edu, https://orcid.org/0000-0003-3187-9719 (EM); georgechen@cmu.edu (GHC); dokyun@bu.edu, https://orcid.org/0000-0002-3186-3349 (DL); mds@cmu.edu, https://orcid.org/0000-0001-6844-7923 (MDS)

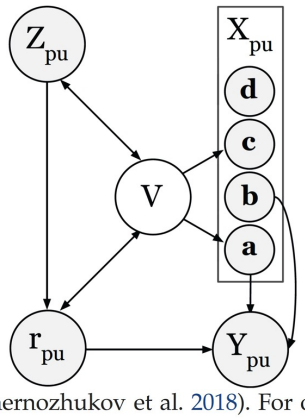
Received: May 31, 2020
Revised: December 13, 2021; May 12, 2022; August 15, 2022
Accepted: August 21, 2022
Published Online in Articles in Advance: May 8, 2023
https://doi.org/10.1287/mnsc.2023.4762

Abstract. Deliberation among individuals online plays a key role in shaping the opinions that drive votes, purchases, donations, and other critical offline behavior. Yet, the determinants of opinion change via persuasion in deliberation online remain largely unexplored. Our research examines the persuasive power of *ethos*—an individual's "reputation"—using a seven-year panel of over a million debates from an argumentation platform containing explicit indicators of successful persuasion. We identify the causal effect of reputation on persuasion by constructing an instrument for reputation from a measure of past debate competition and by controlling for unstructured argument text using neural models of language in the double machine-learning framework. We find that an individual's reputation significantly impacts their persuasion rate above and beyond the validity, strength, and presentation of their arguments. In our setting, we find that having 10 additional reputation points causes a 31% increase in the probability of successful persuasion over the platform average. We also find that the impact of reputation is moderated by characteristics of the argument content, in a manner consistent with heuristic information processing under cognitive overload. We discuss managerial implications for platforms that facilitate deliberative decision making for public and private organizations online.

History: Accepted by Anandhi Bharadwaj, information systems.
Funding: This research was supported in part by the University of Wisconsin Madison Office of the Vice Chancellor for Research and Graduate Education with funding from the Wisconsin Alumni Research Foundation, and by the Carnegie Mellon University Graduate Student Assembly and Provost's Office. Computing support was provided in part by the Social Science Computing Cooperative at the University of Wisconsin Madison.
Supplemental Material: Data and the online appendix is available at https://doi.org/10.1287/mnsc.2023.4762.

Keywords: persuasion • reputation systems • double machine-learning • causal inference with text

DML to evaluate the persuasive power of reputation:
https://pubsonline.informs.org/doi/10.1287/mnsc.2023.4762



learning (Chernozhukov et al. 2018). For our nuisance functions, we use neural networks with rectified linear unit (ReLU) activation functions (Nair and Hinton 2010), which converge at $n^{-1/4}$ rates (Farrell et al. 2018) and, thus, guarantee $\sqrt{n}$-consistent estimates of the causal effect of reputation.

Influence via ethos: On the persuasive power of reputation in deliberation online
E Manzoor, GH Chen, D Lee… - Management …, 2024 - pubsonline.informs.org
… of opinion change via persuasion in deliberation online remain largely unexplored. Our research examines the persuasive power of ethos—an individual's "reputation"—using a seven-…
☆ 保存 99 引用 被引用次数：20 相关文章 所有 7 个版本 Web of Science: 2 »

22

22

Is DML Credible in a Field Setting?

DML vs. RCT to evaluate ad measurement on FB:
https://pubsonline.informs.org/doi/10.1287/mksc.2022.1413

Close Enough? A Large-Scale Exploration of Non-Experimental Approaches to Advertising Measurement

Brett R. Gordon,[a,*] **Robert Moakler,**[b] **Florian Zettelmeyer**[a,c]

[a] Kellogg School of Management, Northwestern University, Evanston, Illinois 60208; [b] Ads Research, Meta, Menlo Park, California 94025; [c] National Bureau of Economic Research, Cambridge, Massachusetts 02138

*Corresponding author

Contact: b-gordon@kellogg.northwestern.edu, https://orcid.org/0000-0001-9081-569X (BRG); rmoakler@meta.com (RM); f-zettelmeyer@kellogg.northwestern.edu (FZ)

Received: January 11, 2022
Revised: May 26, 2022; August 2, 2022
Accepted: August 16, 2022
Published Online in Articles in Advance: November 7, 2022

https://doi.org/10.1287/mksc.2022.1413

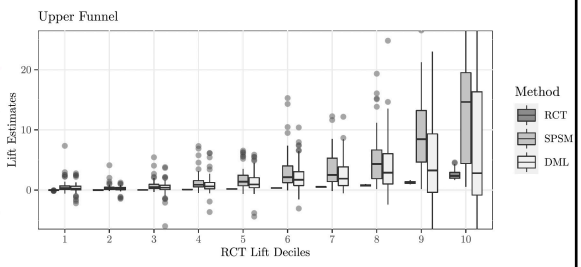
Abstract. Despite their popularity, randomized controlled trials (RCTs) are not always available for the purposes of advertising measurement. Non-experimental data are thus required. However, Facebook and other ad platforms use complex and evolving processes to select ads for users. Therefore, successful non-experimental approaches need to "undo" this selection. We analyze 663 large-scale experiments at Facebook to investigate whether this is possible with the data typically logged at large ad platforms. With access to over 5,000 user-level features, these data are richer than what most advertisers or their measurement partners can access. We investigate how accurately two non-experimental methods—double/debiased machine learning (DML) and stratified propensity score matching (SPSM)—can recover the experimental effects. Although DML performs better than SPSM, neither method performs well, even using flexible deep learning models to implement the propensity and outcome models. The median RCT lifts are 29%, 18%, and 5% for the upper, middle, and lower funnel outcomes, respectively. Using DML (SPSM), the median lift by funnel is 83% (173%), 58% (176%), and 24% (64%), respectively, indicating significant relative measurement errors. We further characterize the circumstances under which each method performs comparatively better. Overall, despite having access to large-scale experiments and rich user-level data, we are unable to reliably estimate an ad campaign's causal effect.

History: Olivier Toubia served as the senior editor for this article.
Funding: To be allowed to access the data required for this paper, B. R. Gordon and F. Zettelmeyer were part-time employees of Facebook with the title of Academic Researchers, employed for three hours per week. R. Moakler is an employee of Meta Platforms, Inc. and owns stock in the company.

Keywords: digital advertising • field experiments • causal inference • observational methods • advertising measurement • double ML

DML and SPSC cannot recover RCT evaluation for advertising measurement.

Question: Which of the assumptions for DML are violated in the FB observational data?



Close enough? a large-scale exploration of non-experimental approaches to advertising measurement
BR Gordon, R Moakler, F Zettelmeyer - Marketing Science, 2023 - pubsonline.informs.org
… we can come "close enough" using the typical data stored on a large ad platform. To answer … These experiments were chosen to be representative of the large-scale experiments …
☆ 保存 99 引用 被引用次数：93 相关文章 所有 13 个版本 Web of Science: 15 »

How Credible is DML in a Field Setting?

Estimating Causal Effects with Double Machine Learning - A Method Evaluation

Jonathan Fuhr[1], Philipp Berens[2], and Dominik Papies[1]

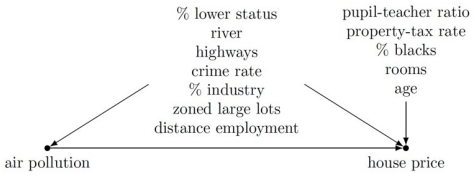
[1] *School of Business and Economics, University of Tübingen, Tübingen, Germany*
[2] *Hertie Institute for AI in Brain Health, University of Tübingen, Tübingen, Germany*

Last edited: March 22, 2024

https://arxiv.org/abs/2403.14385

Abstract

The estimation of causal effects with observational data continues to be a very active research area. In recent years, researchers have developed new frameworks which use machine learning to relax classical assumptions necessary for the estimation of causal effects. In this paper, we review one of the most prominent methods - "double/debiased machine learning" (DML) - and empirically evaluate it by comparing its performance on simulated data relative to more traditional statistical methods, before applying it to real-world data. Our findings indicate that the application of a suitably flexible machine learning algorithm within DML improves the adjustment for various nonlinear confounding relationships. This advantage enables a departure from traditional functional form assumptions typically necessary in causal effect estimation. However, we demonstrate that the method continues to critically depend on standard assumptions about causal structure and identification. When estimating the effects of air pollution on housing prices in our application, we find that DML estimates are consistently larger than estimates of less flexible methods. From our overall results, we provide actionable recommendations for specific choices researchers must make when applying DML in practice.



DML estimates are consistently larger than estimates of less flexible methods (e.g., OLS).

Table 5: Results for the effect of air pollution on housing prices

Method	Effect estimate	Std. error	Effect at mean (%)	MSE(Y)	MSE(W)
OLS (H&R)	-0.0064	0.0011	-7.08	-	-
Simple OLS	-0.0146	0.0011	-16.17	-	-
XGBoost (naive)	-0.0137	0.0013	-15.14	-	-
OLS (raw)	-0.0058	0.0011	-6.47	-	-
OLS (flex)	-0.0071	0.0013	-7.88	-	-
OLS (DML, flex)	-0.0093	0.0006	-10.30	0.5571	1437.38
OLS (DML, raw)	-0.0059	0.0012	-6.58	0.0402	58.91
OLS (DML, H&R)	-0.0064	0.0012	-7.14	0.0375	54.88
GAMs (DML)	-0.0087	0.0015	-9.63	0.0346	42.71
Neural nets (DML)	-0.0081	0.0016	-9.00	0.0349	34.46
Lasso (DML, flex)	-0.0071	0.0015	-7.86	0.0316	33.89
XGBoost (DML)	-0.0070	0.0019	-7.73	0.0295	20.85
Random forests (DML)	-0.0075	0.0018	-8.27	0.0266	19.03

Note: MSE: mean squared error. H&R: covariate specification by Harrison and Rubinfeld (1978). raw: only using untransformed variables. flex: including squares and first-order interactions of all variables.

Hedonic housing prices and the demand for clean air
D Harrison Jr, DL Rubinfeld - Journal of environmental economics and …, 1978 - Elsevier
… the hedonic housing value function, p(h). The p(h) f unction translates a vector of housing attributes at each location into a price … of housing attributes .Q Implicit in this description of the …
☆ 保存 99 引用 被引用次数：2780 相关文章 所有 14 个版本 Web of Science: 1115 »

Estimating Causal Effects with Double Machine Learning--A Method Evaluation
J Fuhr, P Berens, D Papies - arXiv preprint arXiv:2403.14385, 2024 - arxiv.org
… necessary for the estimation of causal effects. In this paper, we review one of the most prominent methods - "double/debiased machine learning" (DML) - and empirically evaluate it by …
☆ 保存 99 引用 被引用次数：10 相关文章 所有 4 个版本 »