Double ML: Causal Inference based on ML

Part I: Introduction to Causal Machine Learning

uai2022, August 1, 2022, Eindhoven

Philipp Bach, Martin Spindler (UHH & Economic AI)

Collaborators: Victor Chernozhukov (MIT), Malte Kurz (TUM)

Motivation for Causal Machine Learning

Motivation for Causal ML

Predictive vs. Causal ML

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Predictive ML

How can we build a good prediction rule, $f(X)$, that uses features X to predict Y?

Example: Customer Churn

> "*How well can we predict whether customers churn?*"

Causal ML

What is the causal effect of a treatment D on an outcome Y?

> "*Why do customer churn?*"
>
> "*How can we retain customers?*"

Motivation for Causal ML

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Typical (research) questions in industry, business and economics:

- What is the effect of the new website (feature) on our sales?
- How much additional revenue did our latest newsletter generate?

Typical research question in academic studies:

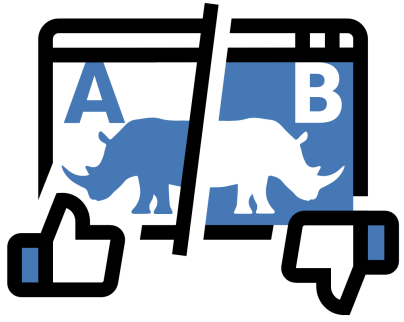
- What is the causal effect of active labor market policies on labor market participation?
- How does a new drug change patients' mortality due to a particular disease?

General Question

What is the causal effect of a treatment D on an outcome Y?

Application: Randomized Experiments

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG



- General: What is the effect of a certain variable D on a relevant outcome variable Y?
- Randomized experiments are a direct way to estimate such effects (assuming they are conducted properly)

Challenges in practice

1. No (pure) A/B-testing / experiments possible -> observational data
2. A/B test suffers from low power
3. Heterogenous treatment effects

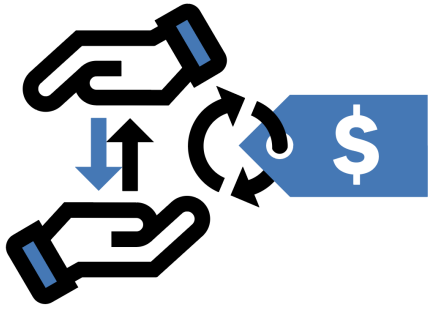
Solution with DoubleML

1. Observational study: Include control variables X which may also impact the variables Y or D
2. Include covariates X that help to predict the outcome Y using ML methods.
3. Detection of complex treatment effect patterns

Example: Price Elasticity of Demand

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Price Elasticity of Demand: How does the price impact sales?



- Absolute change in price (EUR 100) and the resulting absolute change in sales (10 million units) can be difficult to interpret
- Price elasticity of demand: Percentage change in quantity demanded Q when there is a one percent increase in price P

$$E_d = \frac{\Delta Q/Q}{\Delta P/P} = \frac{-10/200}{100/1000} = \frac{-0.05}{0.1} = -0.5$$

Econometric model for estimating the price elasticity θ_0:

$$\log(Q) = \alpha + \theta_0 \log(P) + X'\beta + \varepsilon,$$

where the vector of controls X can be very high-dimensional

Motivation for Causal Machine Learning

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Machine Learning methods usually tailored for prediction
- In science and industry both prediction (stock market, demand, …) and learning of causal relationship is of interest
- Here: Focus on causal inference with machine Learning methods
- Examples for causal inference:
  - Effect of a new website, app design or the latest newsletter
  - Price elasticity of demand
- General: What is the effect of a certain treatment on a relevant outcome variable?

Motivation for Causal Machine Learning

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Challenge I: Identification of causal parameter

Typical problem - Potential endogeneity of the treatment assignment

- Potential sources
  - Optimizing behavior of the individuals with regard to the outcome
  - Simultaneity (price elasticity of demand)
  - Unobserved heterogeneity
  - Omitted variables
  - The treatment assignment is observed rather than randomly assigned
  - Example: Covid vaccination

- Possible Solutions
  - Selection on observable characteristics (controls)
  - Instrumental Variable (IV) estimation
  - ...

Motivation for Causal Machine Learning

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Challenge II: "Big data"

- High-dimensional setting with p (# of variables) even larger than n (# of observations), and / or
- a highly non-linear functional form for the relationship between variables of interest

Solution

- Use ML methods for estimation that ...
  - ... perform regularization, e.g., variable selection
  - ... are able to model non-linearities

Motivation for Causal Machine Learning

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

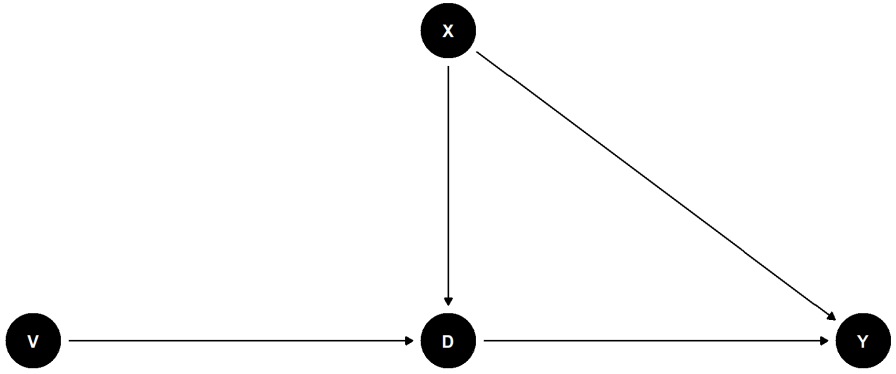
Challenge I & II: Controls are needed for two reasons

Identification:

We need them to ensure D is as good as randomly assigned (exogenous) *conditional on* X. Along these lines, we can think of the confounding equation:

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

$\rightarrow$ Variation in V is quasi-experimental

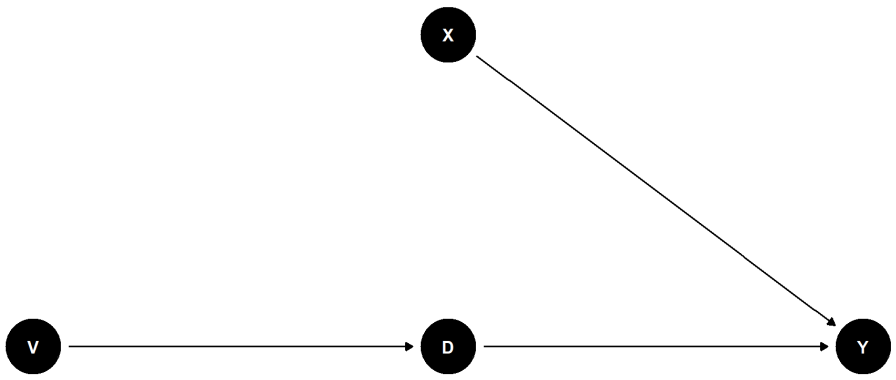


Efficiency:

Some controls may explain part of the variation in the outcome Y

ML methods may deliver more accurate results than OLS or t-tests

$\rightarrow$ The effect can be estimated more accurately



What is Double/Debiased Machine Learning (DML)?

What is Double/Debiased Machine Learning (DML)?

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Double/debiased machine learning (DML) surveyed by Chernozhukov et al. (2018)
- General framework for causal inference and estimation of treatment effects based on machine learning tools using big data
- Combines the strength of machine learning and econometrics
- Our object-oriented implementation DoubleML (in R and Python) provides a general interface for the growing number of models and methods for DML
- Documentation & user guide: https://docs.doubleml.org
- Install the latest release via pip or conda, see installation guide

```
pip install -U DoubleML
```

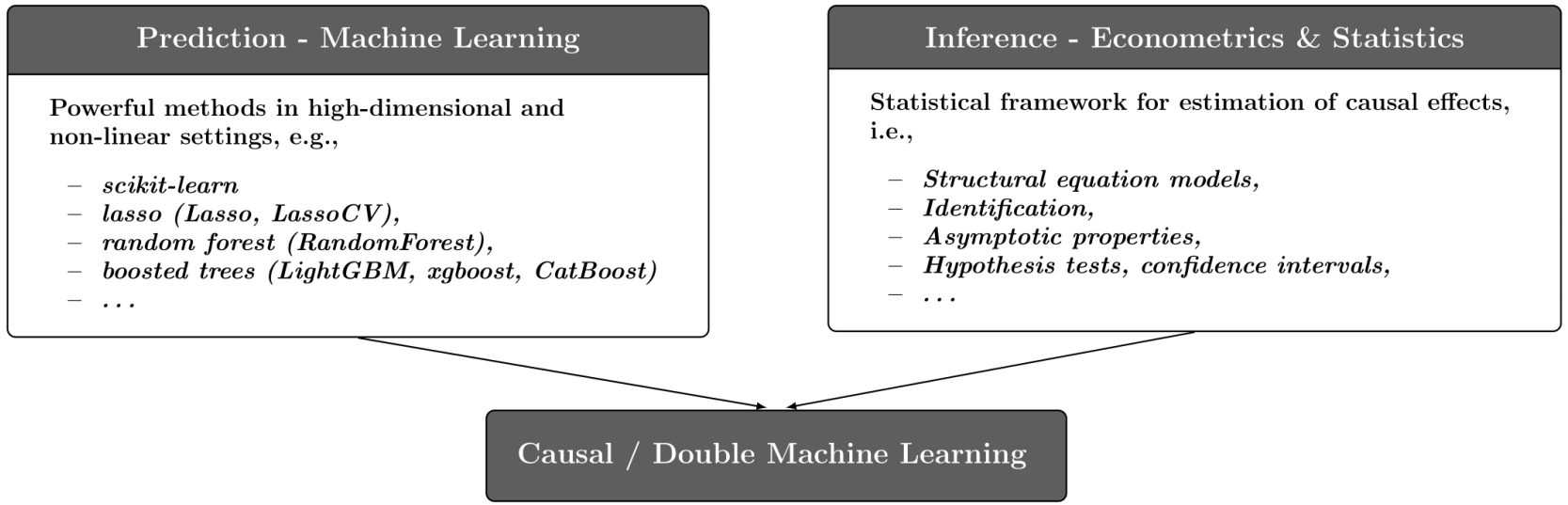
```
conda install -c conda-forge doubleml
```

- Install development version from GitHub https://github.com/DoubleML/doubleml-for-py

What is Double/Debiased Machine Learning (DML)?

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- Exploiting the strengths of two disciplines:



- Result / output from the DML framework:
  - Estimate of the causal effect (with valid confidence intervals → statistical tests for effects)
  - Good statistical properties ($\sqrt{N}$ rate of convergence; unbiased; approximately Gaussian)
  - Multiple treatment effects, heterogeneous treatment effects, ...

A Motivating Example: Basics of Double Machine Learning

Partially Linear Regression

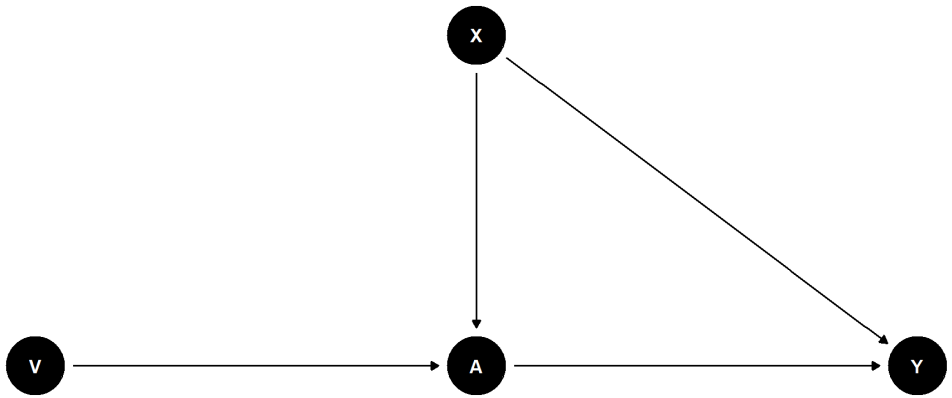
Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Partially linear regression (PLR) model

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

with

- Outcome variable Y
- Policy or treatment variable of interest D
- High-dimensional vector of confounding covariates $X = (X_1, \ldots, X_p)$
- Stochastic errors ζ and V



Partially Linear Regression

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Partially linear regression (PLR) model

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

with

- Outcome variable Y
- Policy or treatment variable of interest D
- High-dimensional vector of confounding covariates $X = (X_1, \ldots, X_p)$
- Stochastic errors ζ and V

Problem of simple "plug-in" approaches: Regularization bias

- If we use an ML model to estimate $\hat{g}$ and simply plug in the predictions $\hat{g}$, the final estimate on θ_0 will not be unbiased and neither be asymptotically normal

Partially Linear Regression

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Illustration of naive approach: App

Example based on Chernozhukov et al. (2018) and

https://docs.doubleml.org/stable/guide/basics.html

App available via GitHub: https://github.com/DoubleML/BasicsDML

Frisch-Waugh-Lovell Theorem

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Solution to regularization bias: Orthogonalization

- Remember the Frisch-Waugh-Lovell (FWL) Theorem in a linear regression model

$$Y = D\theta_0 + X'\beta + \varepsilon$$

- θ_0 can be consistently estimated by partialling out X, i.e,

  1. OLS regression of Y on X: $\tilde{\beta} = (X'X)^{-1}X'Y \rightarrow$ Residuals $\hat{\varepsilon}$
  2. OLS regression of D on X: $\tilde{\gamma} = (X'X)^{-1}X'D \rightarrow$ Residuals $\hat{\zeta}$
  3. Final OLS regression of $\hat{\varepsilon}$ on $\hat{\zeta}$

- Orthogonalization: The idea of the FWL Theorem can be generalized to using ML estimators instead of OLS

Partially Linear Regression

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Illustration of naive approach: App

Example based on Chernozhukov et al. (2018) and

https://docs.doubleml.org/stable/guide/basics.html

App available via GitHub: https://github.com/DoubleML/BasicsDML

The Key Ingredients of Double Machine Learning

The Key Ingredients of DML

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

> 1. Neyman Orthogonality
>
> The inference is based on a score function $\psi(W; \theta, \eta)$ that satisfies
>
> $$E[\psi(W; \theta_0, \eta_0)] = 0,$$
>
> where $W := (Y, D, X, Z)$ and with θ_0 being the unique solution that obeys the Neyman orthogonality condition
>
> $$\partial_\eta \mathbb{E}[\psi(W; \theta_0, \eta]\big|_{\eta=\eta_0} = 0.$$

- ∂_η denotes the pathwise (Gateaux) derivative operator
- Neyman orthogonality ensures that the moment condition identifying θ_0 is insensitive to small pertubations of the nuisance function η around η_0
- Using a Neyman-orthogonal score eliminates the first order biases arising from the replacement of η_0 with a ML estimator $\hat{\eta}_0$

The Key Ingredients of DML

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

> 1. Neyman Orthogonality
>
> The inference is based on a score function $\psi(W; \theta, \eta)$ that satisfies
>
> $$E[\psi(W; \theta_0, \eta_0)] = 0,$$
>
> where $W := (Y, D, X, Z)$ and with θ_0 being the unique solution that obeys the Neyman orthogonality condition
>
> $$\partial_\eta \mathbb{E}[\psi(W; \theta_0, \eta]\big|_{\eta=\eta_0} = 0.$$

- For many models the Neyman orthogonal score functions are linear in θ

$$\psi(W; \theta, \eta) = \psi_a(W; \eta)\theta + \psi_b(W; \eta)$$

- The estimator $\tilde{\theta}_0$ then takes the form

$$\tilde{\theta}_0 = -(\mathbb{E}_N[\psi_a(W; \eta)])^{-1}\mathbb{E}_N[\psi_b(W; \eta)]$$

The Key Ingredients of DML

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

> 1. Neyman Orthogonality
>
> The inference is based on a score function $\psi(W; \theta, \eta)$ that satisfies
>
> $$E[\psi(W; \theta_0, \eta_0)] = 0,$$
>
> where $W := (Y, D, X, Z)$ and with θ_0 being the unique solution that obeys the Neyman orthogonality condition
>
> $$\partial_\eta \mathbb{E}[\psi(W; \theta_0, \eta]\big|_{\eta=\eta_0} = 0.$$

- PLR example: Orthogonality by including the first-stage regression, i.e., the regression relationship of the treatment variable D and the regressors X.
- Orthogonal score function $\psi(\cdot) = (Y - E[Y|X] - \theta(D - E[D|X]))(D - E[D|X])$.

Neyman Orthogonality

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

The two strategies rely on very different moment conditions for identifying and estimating θ_0:

$$\mathbb{E}[\psi(W, \theta_0, \eta_0)] = 0$$

Naive approach

$$\psi(W, \theta_0, \eta) = (Y - D\theta_0 - g_0(X))D$$

Regression adjustment score

$$\eta = g(X),$$
$$\eta_0 = g_0(X),$$

FWL partialling out

$$\psi(W, \theta_0, \eta_0) = ((Y - E[Y|X]) - (D - E[D|X])\theta_0)(D - E[D|X])$$

Neyman-orthogonal score (Frisch-Waugh-Lovell)

$$\eta = (g(X), m(X)),$$
$$\eta_0 = (g_0(X), m_0(X)) = (\mathbb{E}[Y \mid X], \mathbb{E}[D \mid X])$$

Both estimators solve the empirical analog of the moment conditions:

$$\frac{1}{n}\sum_{i=1}^{n} \psi(W_i, \theta, \hat{\eta}_0) = 0,$$

where instead of unknown nuisance functions we plug-in their ML-based (hold-out) estimators

The Key Ingredients of DML

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

> 2. High-Quality Machine Learning Estimators
>
> The nuisance parameters are estimated with high-quality (fast-enough converging) machine learning methods.

- Different structural assumptions on η_0 lead to the use of different machine-learning tools for estimating η_0 (Chernozhukov et al., 2018, Chapter 3)

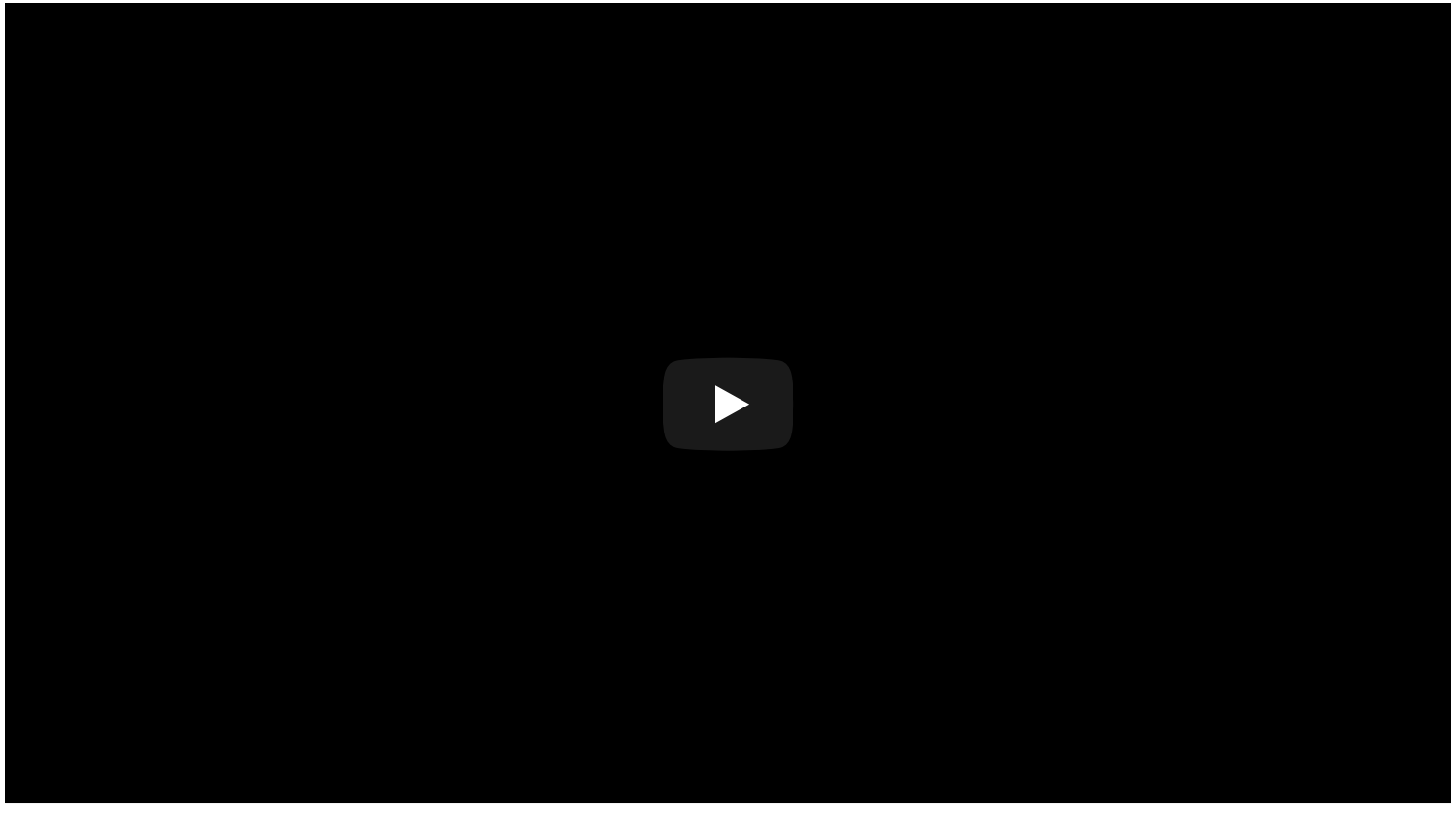
> 3. Sample Splitting
>
> To avoid the biases arising from overfitting, a form of sample splitting is used at the stage of producing the estimator of the main parameter θ_0.

- Cross-fitting performs well empirically (efficiency gain by switching roles)

Key Ingredients of DML

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Illustration of the cross-fitting algorithm



Partially Linear Regression

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Illustration of DML approach: App

Example based on Chernozhukov et al. (2018) and

https://docs.doubleml.org/stable/guide/basics.html

App available via GitHub: https://github.com/DoubleML/BasicsDML

References

References

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Double Machine Learning Approach

- 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: C1-C68, doi:10.1111/ectj.12097.

- Chernozhukov, V., Hansen, C., Kallus, N., Spindler, M., and Syrgkanis, V. (forthcoming), Applied Causal Inference Powered by ML and AI.

DoubleML Package for Python and R

- Bach, P., Chernozhukov, V., Kurz, M. S., and Spindler, M. (2021), DoubleML - An Object-Oriented Implementation of Double Machine Learning in R, arXiv:2103.09603.

- Bach, P., Chernozhukov, V., Kurz, M. S., and Spindler, M. (2022), DoubleML - An Object-Oriented Implementation of Double Machine Learning in Python, Journal of Machine Learning Research, 23(53): 1-6, https://www.jmlr.org/papers/v23/21-0862.html.

Appendix

Examples: Covid Vaccination

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Does the COVID-19 vaccine increase mortality?

Source: Tagesschau.de

Examples: Covid Vaccination

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Does the COVID-19 vaccine increase mortality?