Speaker Notes, Part I

Philipp Bach

UseR! June 20, 2022

Speaker Notes

Speaker notes for the first part of the tutorial *Causal Machine Learning with DoubleML* at the UseR!2022 Conference, June, 2022 by Philipp Bach (philipp.bach@uni-hamburg.de, University of Hamburg), Martin Spindler (martin.spindler@uni-hamburg.de, University of Hamburg), and Oliver Schacht (oliver.schacht@uni-hamburg.de).

Title Slide 1: Causal Machine Learning with DoubleML

- Welcome everybody to our tutorial on Causal Machine Learning with DoubleML.
- Speaker: Philipp Bach, postdoctoral research at the chair of statics with applications in business administration at University of Hamburg.
- Later, my colleagues Martin Spindler who is professor for statistics at the University of Hamburg and Oliver Schacht, who is currently pursuing his PhD in statistics are joining in the hands-on part.
- First of all, we would like to thank the organizators of the UseR! conference for accepting our tutorial proposal and we would like to welcome all participants!

Section Slide: Motivation for Causal Machine Learning

- Let us start with a brief motivation for causal machine learning and first give the participants an impression on what to expect from this tutorial

Slide 3: Motivation for Causal ML

- The slide shows a diagram illustrating causal machine learning.

- Causal machine learning can be considered as an interaction of the research areas of machine learning and causal inference.
- On the one hand side, impressive advances have been made in the literature on artificial intelligence (AI) or machine learning (ML). High-performing algorithms have been developed such as extreme gradient boosting or various neural network architectures, for example. The majority of these innovations have the goal to generate very accurate predictions based on data or features.
- On the other hand, a extensive literature has evolved being concerned with statistical approaches to study causal relationships. Important work has been done by Judea Pearl, Donald Rubin and Guido Imbens, who was co-recipient of the 2021 nobel prize in economics.

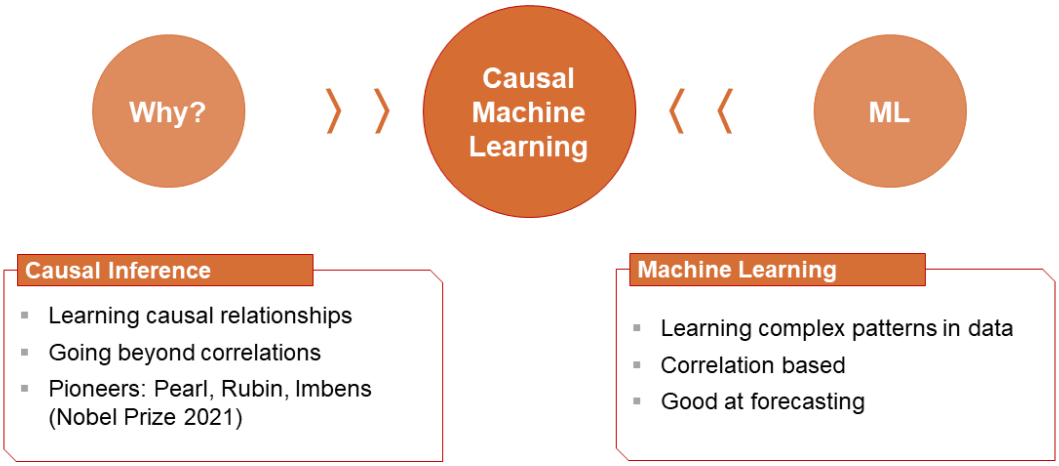


Figure 1: An illustration of Causal Machine Learning based on three circles. The circle in the middle has the title 'Causal Machine Learning'. There are arrows going from the left and right circle pointing at the circle in the center. Below the left circle that carries the title 'Why' a box lists three points: Learning causal relationships, Going beyond correlations, and Pioneers: Pearl, Rubin, Imbens (Nobel Prize 2021). Below the circle on the right which has the title 'ML', three points are listed: Learning complex patterns in the data, correlation based and good at forecasting.

- Purely prediction-focused ML algorithm pursue the goal to generate very precise predictions of an outcome Y based on features X. These prediction rules are based on associations, i.e., they do not really address the underlying causal mechanisms in the data.
- Causal machine learning methods aim to model causal mechanisms and hence go beyond correlations or associations in some sense.

Slide 4: Predictive vs. Causal ML

- To contrast predictive and causal machine learning, we present an example in the context of churn modeling.
- In this example, predictive ML addresses the question: "How well can we predict whether customers churn?"
- Causal ML is used to get insights on why customers churns and what can be done in order to prevent them from churning. Hence, causal ML incorporates the important question "Why?" as well as to think about an intervention, i.e., how to act according to the causal insights.

Slide 5: Motivation for Causal ML

- We list some exemplary research questions for academic use cases
  - What is the **effect** of the **new website (feature)** on our **sales**?
  - Will our new **app design increase** the **time spent** in the app?
  - How much **additional revenue** did our **latest newsletter generate**?
  - Which product would **benefit** most from a **marketing campaign**?
- The general question underlying all these examples is: What is the causal effect of a treatment D on an outcome Y?

Slide 6: Application - Randomized Experiments

- The slide presents an illustration of an AB-Test
-
- AB tests are a typical use case for causal inference and causal machine learning in industry.
- Randomized control trials or randomized experiments are the gold-standard to evaluate the causal effect of a treatment on an outcome.
- Problems in practice:
  1. In many cases pure AB tests are not feasible, e.g., too expensive or risky in the business context
  2. And even if an AB test is feasible, it can lack power
- To address these problems
  1. It's possible to set up a credible and valid causal framework, for example addressing confounding through variables X
  2. Including controls that explain variation in Y can improve the precision of a test

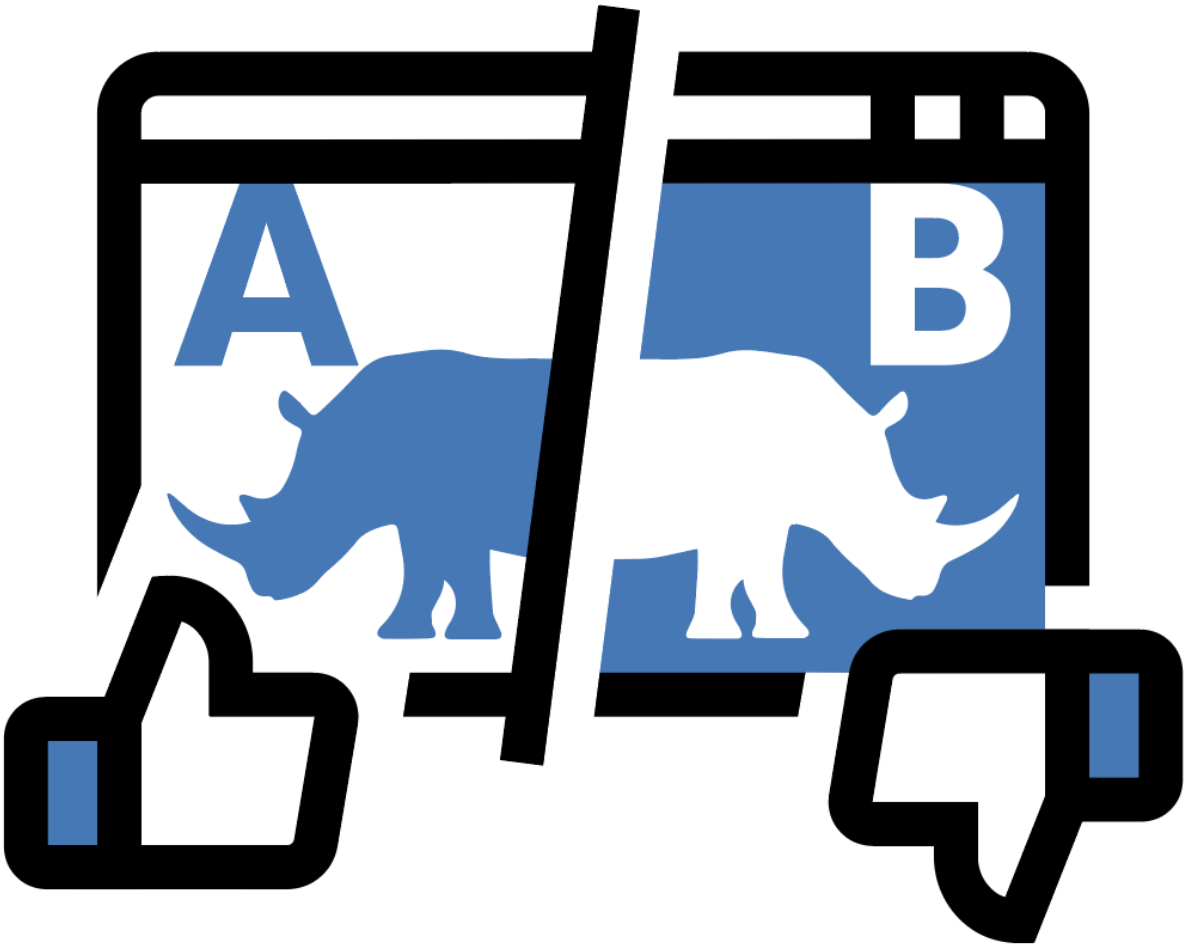


Figure 2: An illustration of AB testing. A stylized browser window shows a double-headed rhino which is a variant of the DoubleML package logo. The screen is divided vertically in two parts. The left part of the screen has the tag 'A' and differs from the right part called 'B' in that the colors are inverted.

Slide 7:

- A second example, which we will revisit later is estimation of price elasticities
- The question here is: "How does the price impact sales?"
- The slide presents an illustration of demand estimation

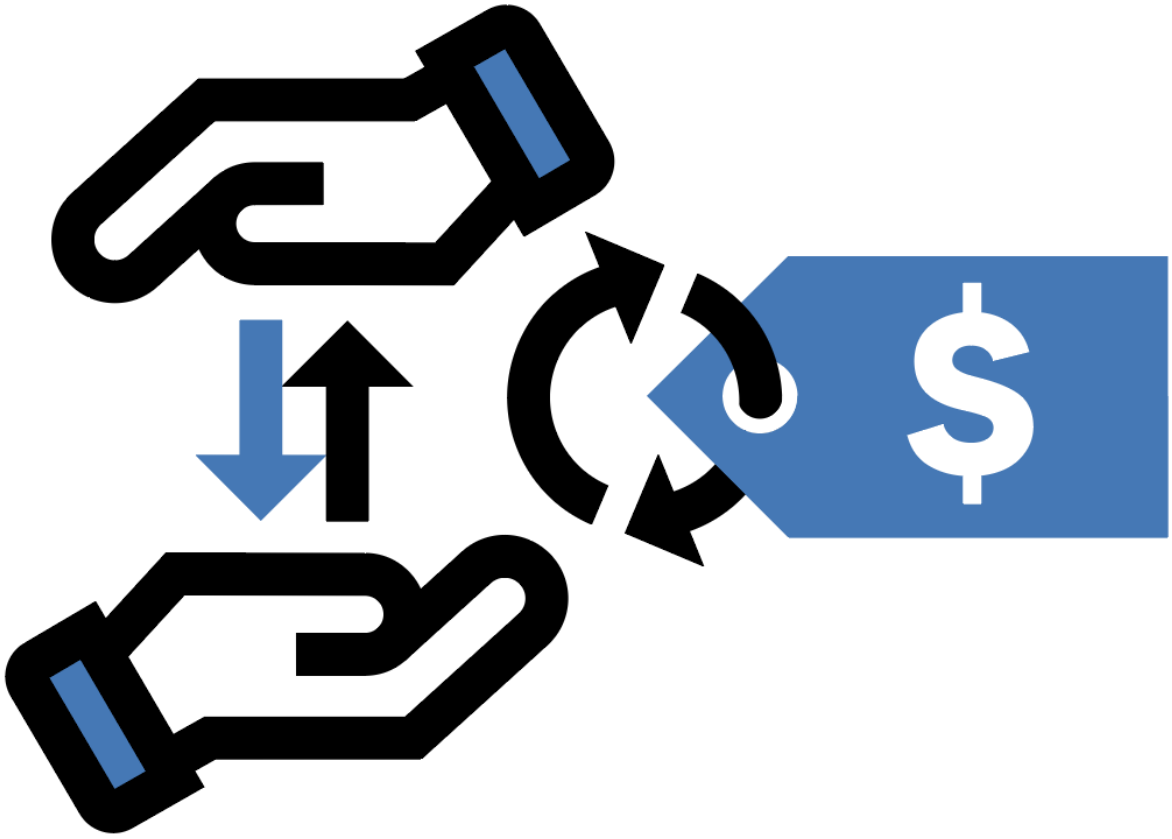


Figure 3: An illustration of demand estimation. On the left hand side two hands are displayed on top of each other. Between the hands there are two arrows showing up and down. On the right hand side, there is a price tag with a dollar sign attached to a circular graph.

- 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
- To estimate the price elasticity that corresponds to the parameter θ_0 we set up a linear regression

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

- There X can be high-dimensional and we might have to use ML methods for estimation

- The underlying causal approach is to sufficiently address the confounding by including the variables X in the regression model

Slide 8: Motivation for Causal Machine Learning

- Let's sum up the motivation for causal ML
- Machine Learning methods usually tailored for **prediction**
- In econometrics 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?

Slide 9: Motivation for Causal Machine Learning

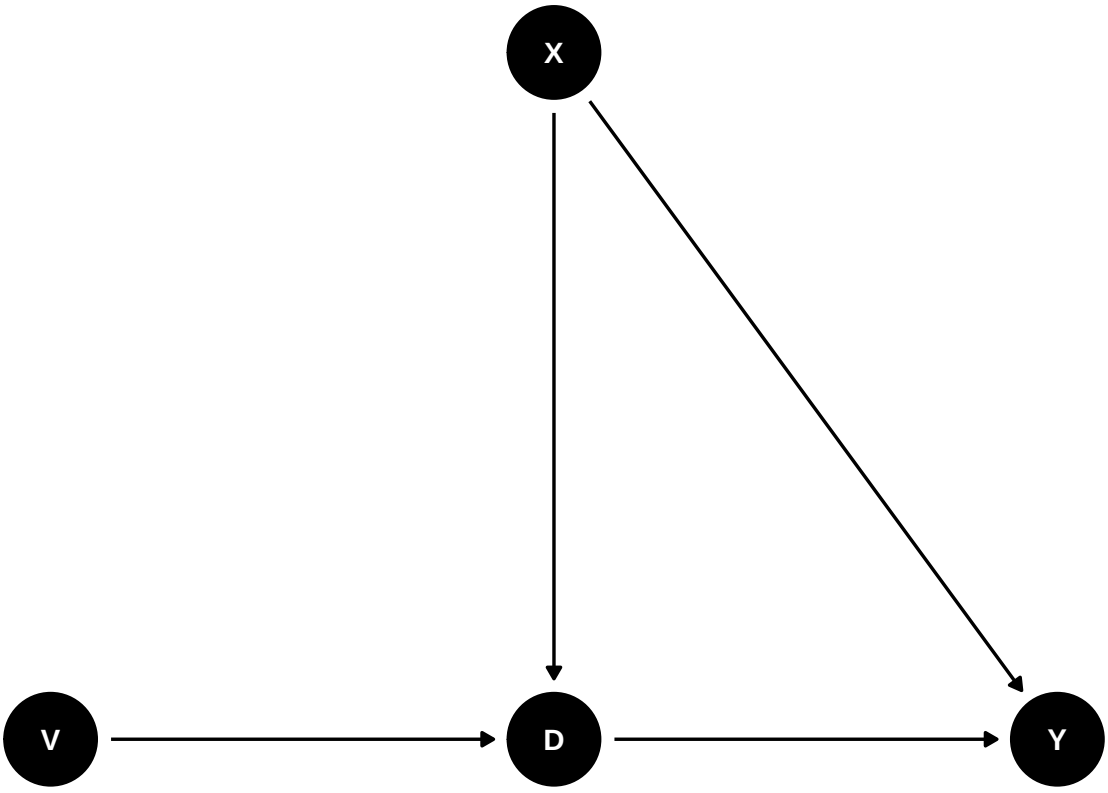
- Causal ML can be considered as addressing two major challenges
- Challenge 1: Identification of causal parameter
- In many cases, we cannot simply use data from an experiment (i.e., a randomly assigned treatment) to assess the causal effect of a treatment on a outcome variable
- **Typical problem** - Potential endogeneity of the treatment assignment.
- The treatment assignment can be confounded or, as it's called in economics, endogenous. There are various reasons that can lead to endogeneity, for example
  - Optimizing behavior of the individuals with regard to the outcome: Individuals self-select into the treatmnet that is optimal for them
  - Simultaneity, for example typical problem for estimation of price elasticity of demand: Several agents make decisions simultaneously and interactively
  - There might be unobserved heterogeneity or important variables being unobserved that drive the treatment assignment
  - As an example, consider an evaluation of the Covid vaccine in 2021. Without going into too much detail: A simple comparison of those who received the vaccine to those who didn't might not uncover the true effect of the vaccine on mortality. In many countries, access to the vaccine has been granted according to risk factors. These factors need to be taken into account when evaluating such a trial.
- To address this challenge, we need to set up a causal model that is able to identify the true causal effect of the treatment on the outcome.
  - Such an approach might involve to account for confounding factors that affect treatment status and the outcome. This approach is called "selection on observables"
  - In other cases, we might have to exploit instrumental variables or make use of more evolved approaches, such as difference-in-differences or synthetic controls.

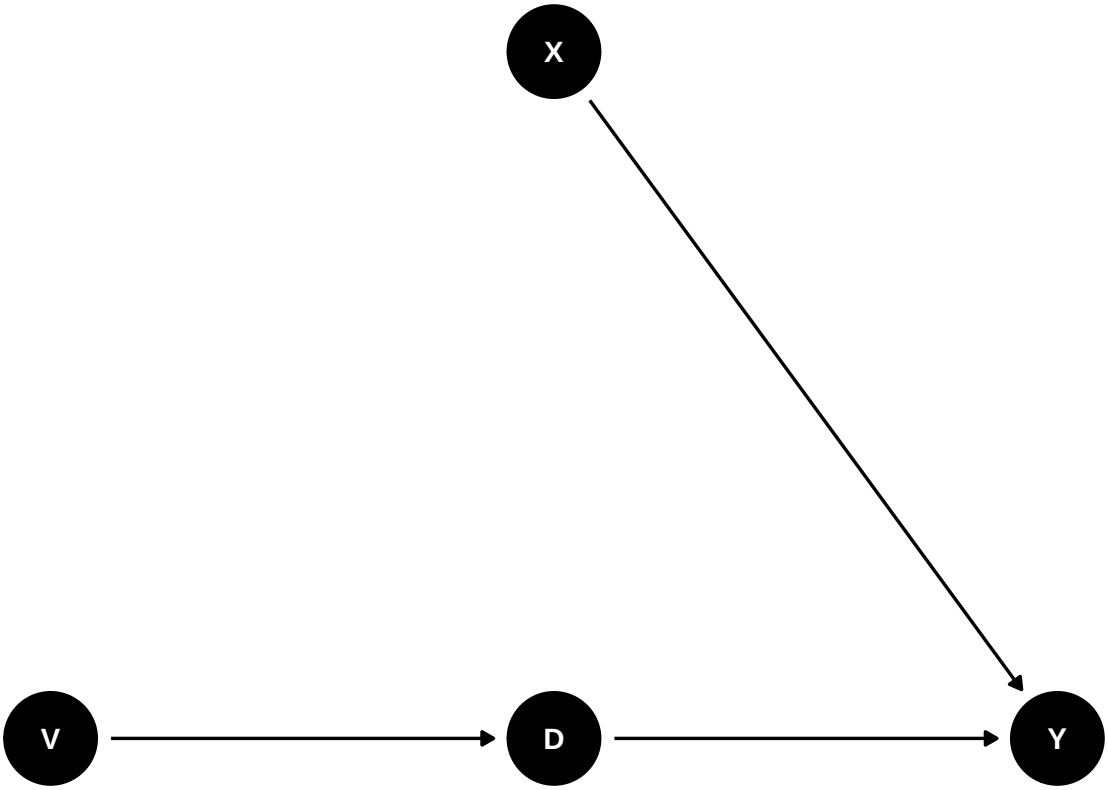
Slide 10: Motivation for Causal Machine Learning

- Challenge 2: "Big Data"
- The second challenge arises from modeling the underlying regression relationships, which might be high-dimensional or very complex
- This challenge can be addressed by using state-of-the-art ML methods that perform regularization, for example variable selection, or that are able to model nonlinearities, like tree-based methods or neural networks

Slide 11: Motivation for Causal Machine Learning

- To sum up, control variables might be included for two reasons
- First, we might to include the variables for identification, for example in a selection-on-observables approach
- Second, we might want to use them to improve precision of our estimator
- Both these approaches can be represented in terms of Directed Acyclical Graphs (DAGs)





Slide 12: Title Slide: What is Double/Debiased Machine Learning (DML)?

Slide 13: What is Double/Debiased Machine Learning (DML)?

- Double/Debiased machine learning (DML) was introduced in a study by Chernozhukov et al. (2018)
- It provides a general framework for causal inference and estimation of treatment effects that is based on ML tools
- DML combines the strengths of ML and causal inference / econometrics
- The object-oriented implementation of the DoubleML package provides a general interface for the increasing number of DML models and methods
- Documentation and a detailed user guide are available via https://docs.doubleml.org.
- The R package can be installed from CRAN.

Slide 14: What is Double/Debiased Machine Learning (DML)?

- To give you an idea on what to expect from double machine learning and our package DoubleML, we'll consider a diagram shown on this slide

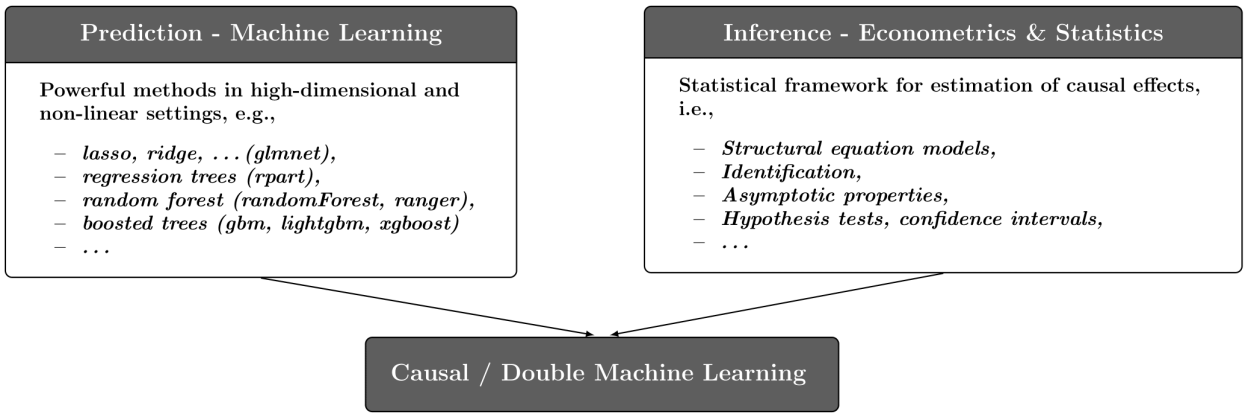


Figure 4: A diagram with three boxes. On top, there are two boxes. On the left, a box with title PREDICTION - MACHINE LEARNING; on the right a box with title INFERENCE - ECONOMETRICS AND STATISTICS. In the first box (prediction) it is written: Powerful methods in high-dimensional and non-linear settings, for example, lasso, ridge, regression trees, random forests and boosted trees. In the second box (inference) it is written: Statistical framework for estimation of causal effects, this is, structural equation models, identification, asymptotic theory, hypothesis tests and confidence intervals.

- The DML approach will output an estimate of the causal effect together with standard errors and confidence intervals, i.e., as a user of DoubleML you will be able to not only estimate the causal parameter of interest, but also to assess whether you can reject a null hypothesis on this quantity at a prespecified significance level
- Moreover, in the DML paper by Chernozhukov et al. (2018), they show that the DML estimator satisifies attractive theoretical properties, such as root-n rate of convergence and approximate normality
- It is also possible to make use of additional results, for example for simultaneous inference and heterogenous treatment effects. The literature on DML extensions is rapidly growing.

Slide 15: Title Slide: A Motivating Example - Basics of Double Machine Learning

Slide 16: Partially linear regression

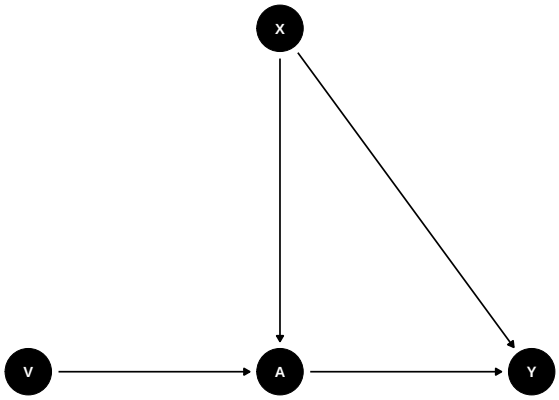
- Let us briefly consider an example that is based on a partially linear regression model. In this example, we will point at the challenges for causal machine learnings and on how to address them in the DML approach.
- Consider the partially linear regression model that is defined by the equations

$$Y = D\theta_0 + g_0(X) + \zeta,$$

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

with $\mathbb{E}[\zeta|D, X] = 0$, $\mathbb{E}[V|X] = 0$ and outcome variable Y, policy or treatment variable of interest D, High-dimensional vector of confounding covariates $X = (X_1, \ldots, X_p)$, and stochastic errors ζ and V.

- The causal model can be represented in a DAG



- Suppose now, we want to use machine learning methods in this model, for example, we might want to estimate the function $g_0(X)$ using an ML learner like a random forest.

Slide 17: Partially linear regression

- The problem with such a naive plug-in approach is that the resulting estimator on θ_0 will generally be biased
- This bias arises due to the regularization that is introduced by virtually any kind of ML method

Slide 18: Partially linear regression

- We'll illustrate the failure of the naive plug-in approach in a simulated data example that is based on Chernozhukov et al. (2018) and

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

- The illustration is based on a shiny app that is available from the GitHub repository https://github.com/DoubleML/BasicsDML
- To run the shiny app locally, use the following commands

```
install.packages("shiny")
shiny::runGitHub("BasicsDML", "DoubleML")
```

- The shiny app shows the empirical distribution of an ML-absed estimator for θ_0 in the partially linear regression model.
- In case of a non-orthogonal score (unchecked checkbox "Orthogonal learner"), the empirical distribution will generally not be normal as well as not centered around the true value of θ_0: The regularization leads to a substantial bias of the naive estimators.

Slide 19: Frisch-Waugh-Lovell Theorem

- The general idea to overcome the regularization bias is to base the estimation on a principle called "orthogonality"
- As an example, consider the Frisch-Waugh-Lovell theorem. The corresponding estimator is also known as "partialling out" ordinary least squares.
- Consider a linear regression model

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

- The coefficient θ_0 can be consistently estimated based on a partialling-out approach, i.e.,
  1. We first run an OLS regression of Y on X
  2. Then we run an OLS regression of D on X
  3. We regress the residuals from step 1. on those obtained from step 2. and obtain a consistent estimator for θ_0
- The idea of the FWL theorem can be generalized to so-called "orthogonality", i.e., with an orthogonal score we are able to replace the OLS regression steps in the FWL approach by ML methods

Slide 20: Partially Linear Regression

- If we tick the checkbox "Orthogonal Score", we see that the empirical distribution is still not yet close to a standard-normal distribution. The reason for this is that orthogonality is a necessary but not yet sufficient ingredient to valid causal machine learning.
- We'll learn more about the other key ingredients in a minute. For now, tick the checkboxes for "Sample Splitting" and choose the option "Tuned" for the "High Quality ML" dropdown menu. You'll see that the corresponding estimator, i.e., the double machine learning estimator, now has an empirical distribution that is very similar to a standard normal as well as being centered around the true value of θ_0.

Slide 21: Title Slide: The Key Ingredients of Double Machine Learning

Slide 22: The Key Ingredients to DML

- As explained before, the DML approach is a general approach to estimate a causal parameter θ_0 based on an orthogonal score function.
- Identification of the parameter of interest, θ_0 is based on the solution of a moment equation with a score function ψ, data W and a nuisance term η, which has population value η_0.
- The DML approach can be formulated in terms of three key ingredients:
  - Neyman Orthogonality,
  - Accurate ML methods, and
  - Sample splitting.
- The first ingredient, Neyman orthogonality, is a property on the score function ψ. It says that the moment function that is used for identification of the causal parameter is insensitive to small errors in the nuisance part η around the true value η_0. This implies that estimation is somehow *immunized* against the regularization bias, that arises when we replace the true value η_0 by its ML estimate $\hat{\eta}_0$.

Slide 23: The Key Ingredients of DML

- In many cases, the Neyman orthogonal score functions are linear in θ, i.e., they can be expressed as

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

which facilitates the computation of θ in practice.

Slide 24: The Key Ingredients of DML

- In the PLR example from before, we can achieve Neyman orthogonality by including the first-stage regression, i.e., the regression relationship of the treatment variable D and the regressors X.
- This leads to the score function $\psi(\cdot) = (Y - g(x) - \theta D)(D - m(X))$. Now, we have a nuisance parameter η that has two components: The function g from the main equation of the PLR and the function m from the first-stage regression.

Slide 25: The Key Ingredients to DML

- The second ingredient to the DML framework is a requirement on the quality of the ML methods in use. In theoretical terms, it is required that the estimation procedures possess fast-enough rates of convergence. These rates are available for many learners.
- In an application, the choice of the ML method will generally depend on the structural assumptions we impose. For example, when we encounter a sparse setting, we might use $l1$ penalized estimators such as the lasso.
- The third ingredient is the use of sample splitting which is necessary to avoid biases due to overfitting.
- In the DML algorithm, we fit the learners on a *training sample* and generate predictions on a *test sample*. We can swap the roles of the samples, which is then called *cross-fitting*. Cross-fitting makes it possible to use all observations in the sample and, hence, achieve efficiency gains.
- The predictions from the cross-fitting procedure are then plugged in into the score function which is then solved for θ_0.

Slide 26: The Double Machine Learning Framework

- The slide shows an illustration of the cross-fitting algorithm which is available via https://youtu.be/BMAr27rp4uA

  alt-text: First, a causal model is defined together with an orthogonal score function. Next, the data is split into K partitions. The learners are fit on the *training samples* and generate predictions on the *test samples*. The predictions are plugged into the score function which is then solved. As a result, we obtain the coefficient estimate together with the standard error. The result in Chernozhukov et al. (2018) says: "Under regularity assumptions, the double machine learner $\tilde{\theta}_0$ is asymptotically normally distributed."

Slide 27: Partially Linear Regression

- We return to the shiny app and can play around with the key ingredients by (un)ticking the checkboxes and selecting the quality of the learner via the dropdown menu.

Slide 28: Title Slide: References

Slide 29: References

- That's it for the first part. On the last slides I added the main references.
- Thank you very much!