Causal Machine Learning with DoubleML

Introduction to the R Package DoubleML

UseR!2022, June 20, 2022, online

Philipp Bach, Martin Spindler, Oliver Schacht (Uni Hamburg)

Introduction to DoubleML

Building Principles

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Key ingredient and Implementation

- Orthogonal Score
  - Object-oriented implementation with R6
  - Exploit common structure being centered around a (linear) score function $\psi(\cdot)$
- High-quality ML
  - State-of-the-art ML prediction and tuning methods
  - Provided by `mlr3` ecosystem
- Sample Splitting
  - Built-in resampling schemes of `mlr3`

Dependency

R6

mlr

mlr

Dependencies and Installation

DoubleML package dependencies

- `mlr3`
- `mlr3learners`
- `mlr3tuning`
- `R6`
- `data.table`

Why an Object-Orientated Implementation?

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

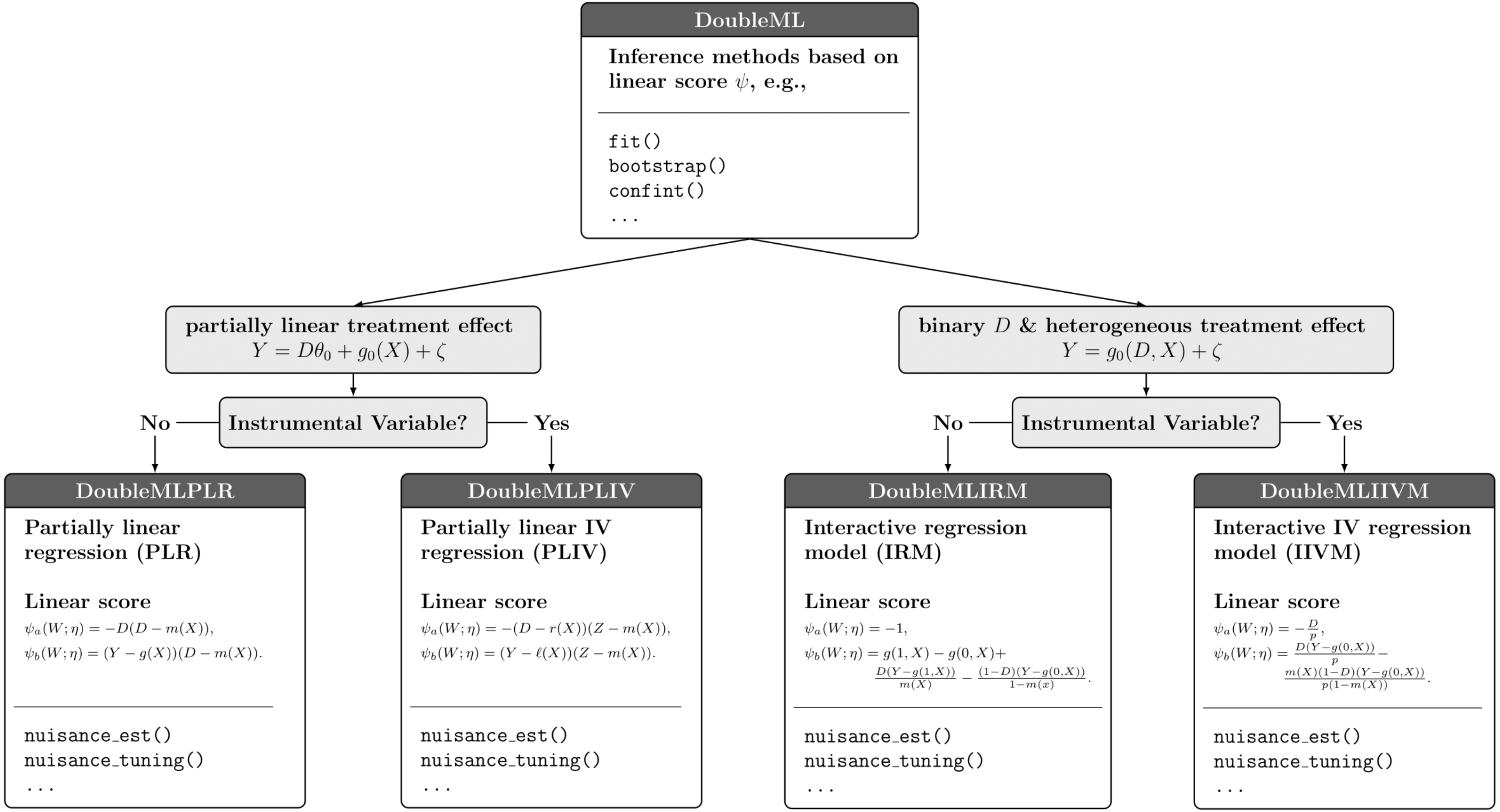
- Given the components $\psi^a(\cdot)$ & $\psi^b(\cdot)$ of a linear Neyman orthogonal score function $\psi(\cdot)$, a general implementation is possible for
  - The estimation of the orthogonal parameters
  - The computation of the score $\psi(W;\theta,\eta)$
  - The estimation of standard errors
  - The computation of confidence intervals
  - A multiplier bootstrap procedure for simultaneous inference
- The sample splitting can be implemented in general as well

$\longrightarrow$ Implemented in the abstract base class `DoubleML`

- The score components and the estimation of the nuisance models have to be implemented model-specifically

$\longrightarrow$ Implemented in model-specific classes inherited from `DoubleML`

Class Structure and Causal Models



Advantages of the Object-Orientation

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- `DoubleML` gives the user a high flexibility with regard to the specification of DML models:
    - Choice of ML methods for approximating the nuisance functions
    - Different resampling schemes (repeated cross-fitting)
    - DML algorithms DML1 and DML2
    - Different Neyman orthogonal score functions

- `DoubleML` can be easily extended
    - New model classes with appropriate Neyman orthogonal score function can be inherited from `DoubleML`
    - The package features `callables` as score functions which makes it easy to extend existing model classes
    - The resampling schemes are customizable in a flexible way

Getting started with DoubleML!

Installation

- Latest *CRAN* release

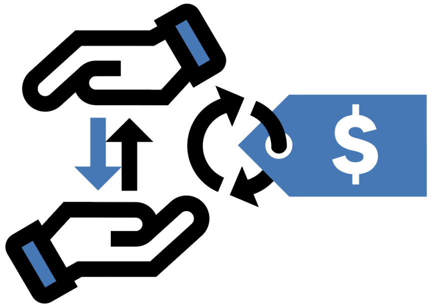
```r
install.packages("DoubleML")
```

- Development version

```r
remotes::install_github("DoubleML/doubleml-for-r")
```

- See the Getting Started page of the tutorial website for more information on prerequisites.

Data Example: Demand Estimation



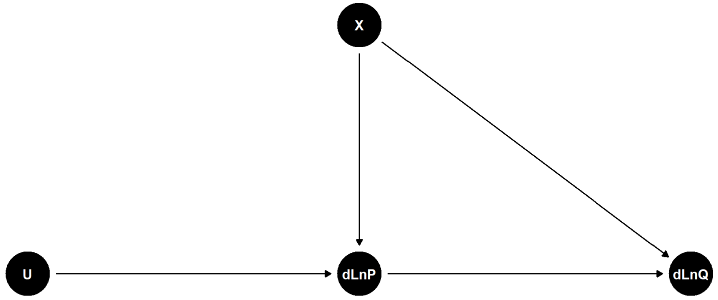
Causal Problem

- Price elasticity of demand: What is the effect of a price change, $dLnP$, on demanded quantity, $dLnQ$?
- Observational study: Flexibly adjust for confounding variables X, e.g. product characteristics

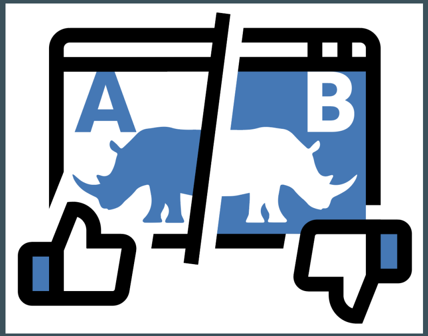
Data Source

- Data example based on a blogpost by Lars Roemheld (Roemheld, 2021)
- Original real data set publicly available via kaggle, preprocessing notebook available online

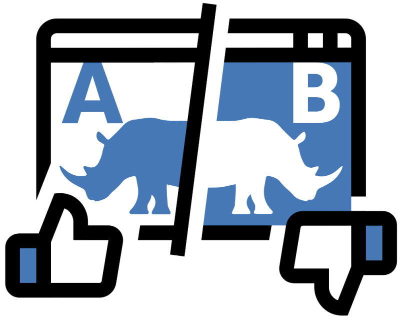
Causal Diagram (DAG)



Hands on! Interactive Breakout Sessions



Data Example: A/B Testing



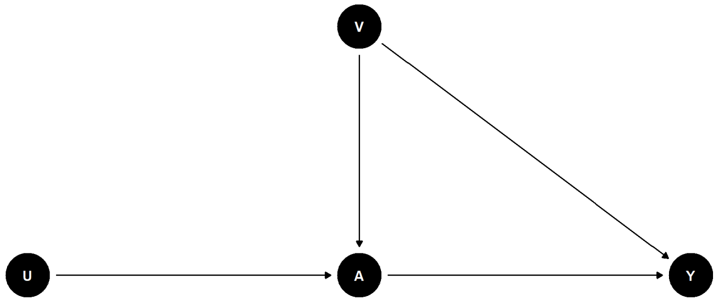
Causal Problem

- Online shop: What is the effect of a new ad design A on sales Y (in $\$100$)?
- Observational study: Necessary to adjust for confounding variables V

Data Source

- Data example based on a randomly chosen DGP created for the 2019 ACIC Data Challenge.

Causal Diagram (DAG)



Online Resources

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

- The notebook is organized according to the DoubleML Workflow
- Extensive User Guide available via docs.doubleml.org
- Documentation for the R Package DoubleML available via docs.doubleml.org/r/stable/
- R vignette, Bach et al. (2021) available via arxiv

Quickstart in R6

- A short introduction to the R6 packages is available here.
- To create a new instance of a class, call the `$new()` method.

```r
# Example create a backend (class DoubleMLData)
library(DoubleML)
df = make_plr_CCDDHNR2018(return_type = "data.table")
obj_dml_data = DoubleMLData$new(df,
                                y_col = "y",
                                d_cols = "d")
```

Quickstart to R6

- Call methods and access fields

```
obj_dml_data$n_obs
```

```
## [1] 500
```

```
obj_dml_data$print()
```

```
## ================= DoubleMLData Object ==================
##
##
## ------------------ Data summary      ------------------
## Outcome variable: y
## Treatment variable(s): d
## Covariates: X1, X2, X3, X4, X5, X6, X7, X8, X9, X10, X11, X12, X13, X14, X15, X16, X17, X18, X19, X20
## Instrument(s):
## No. Observations: 500
```

Quickstart to R6

Quickstart: Creating learners in mlr3

- Install and load `mlr3` package

```r
install.packages("mlr3")
library(mlr3)
```

- Create a learner

```r
lm_learner = LearnerRegrLM$new()
```

```r
lm_learner = lrn("regr.lm")
lm_learner
```

```
## <LearnerRegrLM:regr.lm>
## * Model: -
## * Parameters: list()
## * Packages: mlr3, mlr3learners, stats
## * Predict Type: response
## * Feature types: logical, integer, numeric, factor, character
## * Properties: loglik, weights
```

Thank you UseR!2022

Acknowledgements

- We'd like to thank Malte Kurz and Victor Chernozhukov as well as our research assistants Mehmet Korkmaz, Anzony Quispe, Gangli Tan and Joshua Falke for their support during the preparation of the tutorial.

- We'd like to thank the organizors of the 2019 ACIC Data Challenge for sharing the DGPs and granting us the right to use and distribute data sets.

References

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., 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.