DoubleML Tutorial

Zachary Goldstein

2023-05-12

Introduction

This is a tutorial for how to use the DoubleML package in R. The package implements Double/Debiased Machine Learning for estimating causal effects.

Load Libraries

```
library(DoubleML)
library(tidyverse)
library(causaldata)
```

Example Data

For this tutorial, we'll be using data from the Current Population Survey.

The data lets us study the effect of participation in a job-training program on future wages.

The data is observational, it is not from an experiment and we can not assume that program participation is independent of the potential outcomes.

Besides the main treatment and outcome variables, we have data on some covariates that we hypothesize to be related to participation and wages, including pre-program wages, race, age, educational attainment, and marital status.

For the purposes of this tutorial, we are going to assume that the observed covariates are sufficient to control for confounding. (comparisons with randomized experiment data have shown this may not actually be the case)

```
df = cps_mixtape
df %>% head()
```

```
## # A tibble: 6 x 11
##   data_id treat   age  educ black  hisp  marr nodegree   re74   re75   re78
##   <chr>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>    <dbl>  <dbl>  <dbl>  <dbl>
## 1 CPS1        0    45    11     0     0     1        1 21517. 25244. 25565.
## 2 CPS1        0    21    14     0     0     0        0  3176.  5853. 13496.
## 3 CPS1        0    38    12     0     0     1        0 23039. 25131. 25565.
## 4 CPS1        0    48     6     0     0     1        1 24994. 25244. 25565.
## 5 CPS1        0    18     8     0     0     1        1  1669. 10728.  9861.
## 6 CPS1        0    22    11     0     0     1        1 16366. 18449. 25565.
```

Partially Linear Model

Let's start with a relatively simple option for Double Machine Learning, the partially linear model. A partially linear model assumes a linear relationship between the treatment and outcome, but makes no such assumptions about the relationships between the covariates and the outcome. We can use non-parametric machine learning methods of our choice to model the relationship between the covariates and the outcome.

The model is as follows:

$$Y = D\theta_0 + g_0(X) + \zeta D = m_0(X) + VE[\zeta|D, X] = 0E[V|X] = 0$$

Y is the continuous outcome variable **re78**, real earnings in 1978. D is the binary treatment variable **treat** corresponding to job training program participation.

$$/theta$$

is