Seminar Data Science for Economics

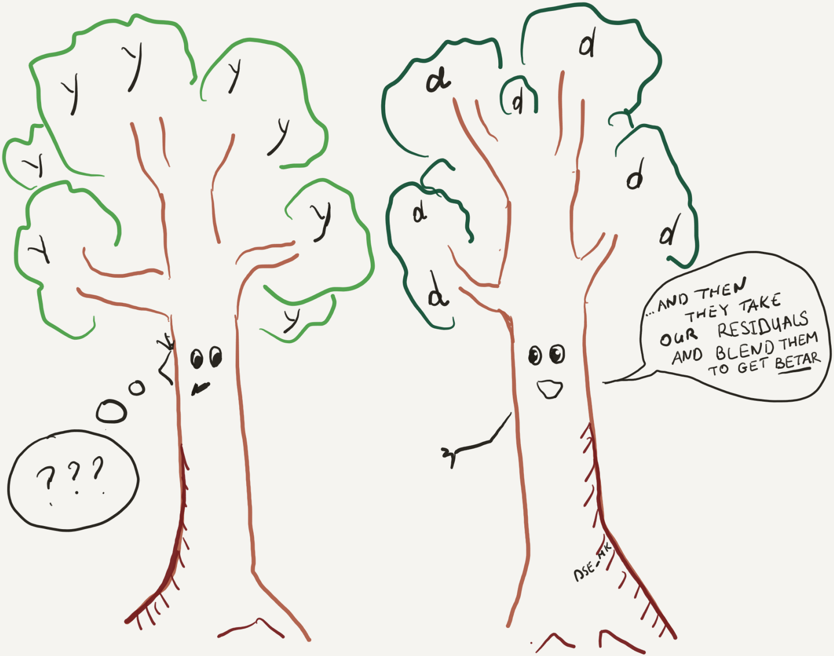
MSc. Economics program

Madina Kurmangaliyeva
m.kurmangaliyeva@uvt.nl

Spring 2020

Tilburg University

Double Machine Learning (Chernozhukov et al., 2018)


? ? ?
...AND THEN THEY TAKE OUR RESIDUALS AND BLEND THEM TO GET BETAR
DSE_MK

Partially Linear Regression setting

Outcome equation:

$$y_i = \beta_0 d_i + g_0(Z_i) + u_i, \quad E(u_i|Z_i, d_i) = 0 \tag{1}$$

Treatment equation:

$$d_i = m_0(Z_i) + v_i, \quad E(v_i|Z_i) = 0 \tag{2}$$

- y_i - outcome variable
- d_i - treatment variable
- $Z_i = (z_{1,i}, \dots, z_{p,i})$ - vector of controls
- $g_0(.)$ and $m_0(.)$ are some functions

We are interested in β_0, which is the treatment effect

(Naive) Double Machine Learning

1. Use ML* to predict y from $z \implies$ get $y - \hat{g}_0(Z)$
2. Use ML* to predict d from $z \implies$ get $\hat{v} = d - \hat{m}_0(Z)$
3. Regress residuals on residuals: $y - \hat{y}$ on $d - \hat{d}$ to get treatment effect

ML* can be Ridge, Lasso, Neural Network, Decision Trees, etc.

The estimator:

$$\hat{\beta}_0 = \left(\frac{1}{n} \sum_{i \in 1}^{n} \hat{v}_i d_i \right)^{-1} \frac{1}{n} \sum_{i \in 1}^{n} \hat{v}_i \left(y_i - \hat{g}_0(Z_i) \right) \tag{3}$$

It looks a lot like an IV estimator, where the instrument is the prediction error.

Overfitting bias

Problem: Naive DML is biased

- We used the same sample for predicting functional forms of $g_0(Z)$ and $m_0(Z)$ AND for estimating $\hat{\beta}$
- Theory shows it leads to overfitting bias because $cov(v_i, \hat{g}_0(Z_i) - g_0(Z_i)) \neq 0$
- This term – $cov(v_i, \hat{g}_0(Z_i) - g_0(Z_i))$ – affects asymptotic bias of $\hat{\beta}$ (see discussion on page 5 of the paper.
- Remember that we saw the downward bias already in tutorial 2 when we used cross-validated Lasso? We also see it in tutorial 3

Solution: split sample into **training** and **estimation** sets to ensure that $cov(v_i, \hat{g}_0(Z_i) - g_0(Z_i)) = 0$

Cross-fitted DML estimator

1. Randomly split the sample into halves: I and I^c
2. Use ML to predict y from Z on $I^c \implies$ get the residuals $y - \hat{g}_0(Z)$ for observations in I
3. Use ML to predict d from Z on $I^c \implies$ get the residuals $\hat{v}$ for observations in I
4. Regress residuals on residuals: $y - \hat{y}$ on $d - \hat{d}$ to get the estimate of $\hat{\beta}_0^{(1)}$ for sample I:

$$\hat{\beta}_0^{(1)} = \left(\frac{1}{0.5n} \sum_{i \in I} \hat{v}_i d_i \right)^{-1} \frac{1}{0.5n} \sum_{i \in I} \hat{v}_i \left(y_i - \hat{g}_0(Z_i) \right) \tag{4}$$

Cross-fitted DML estimator (cont.d)

5. Repeat Steps 2-4 but now using set I for prediction and set I^c for estimation. Get $\hat{\beta}_0^{(2)}$ for sample I^c
6. The cross-fitted DML estimator:

$$\hat{\beta} = 0.5\left(\hat{\beta}_0^{(1)} + \hat{\beta}_0^{(2)}\right) \tag{5}$$

In principle, you can make K splits of data... just like K-fold CV, it is K-fold cross-fitting.

For the asymptotic variance formula, see pages 4, 26-27 of the paper

DML vs DS

The idea of DML is very similar to Double Selection. BUT:

- In comparison to DS, we do not use Lasso to select variables, but rather use ML to remove nuisance parameters from y and d. And then, we directly use the partialling out estimator.
- We need to cross-fit to overcome regularization biases
- DML approach can be generalized for the setting with heterogeneous treatment effects, i.e.:

$$y_i = \underbrace{g_0(d_i, Z_i)}_{\text{TE not separable}} + u_i, \quad E(u_i|Z_i, d_i) = 0 \tag{6}$$