Prediction models

Seminar Data Science for Economics

Madina Kurmangaliyeva
m.kurmangaliyeva@uvt.nl

Spring 2021

Tilburg University

Partialling-out via Post-Lasso

The Partialling-out via Post-Lasso procedure in three steps:

1. Step 1:
    - 1.1 Run Lasso for y on Z to select a subset $Z^{\star,y} \in Z$ that best predicts y
    - 1.2 Run OLS for y on the selected $Z^{\star,y} \in Z \Longrightarrow$ Get residuals $\dot{y}$
2. Step 2:
    - 2.1 Run Lasso for d on Z to select a subset $Z^{\star,d} \in Z$ that best predicts d
    - 2.2 Run OLS for d on the selected $Z^{\star,d} \in Z \Longrightarrow$ Get residuals $\dot{d}$
3. Step 3:
    - 3.1 Run an OLS for $\dot{y}$ on $\dot{d}$

WHY POST-LASSO?

The Partialling-out via Post-Lasso procedure in three steps:

1. Step 1:
   1.1 Run Lasso for y on Z to select a subset $Z^{\star,y} \in Z$ that best predicts y
   1.2 Run OLS for y on the selected $Z^{\star,y} \in Z \Longrightarrow$ Get residuals $\dot{y}$
2. Step 2:
   2.1 Run Lasso for d on Z to select a subset $Z^{\star,d} \in Z$ that best predicts d
   2.2 Run OLS for d on the selected $Z^{\star,d} \in Z \Longrightarrow$ Get residuals $\dot{d}$
3. Step 3:
   3.1 Run an OLS for $\dot{y}$ on $\dot{d}$

Partialling-out via ~~Post-Lasso~~

WHY POST-LASSO?

The Partialling-out via Post-Lasso procedure in three steps:

1. Step 1:
   - 1.1 Run Lasso for y on Z to select a subset $Z^{\star,y} \in Z$ that best predicts y
   - 1.2 Run OLS for y on the selected $Z^{\star,y} \in Z \Longrightarrow$ Get residuals $\dot{y}$
2. Step 2:
   - 2.1 Run Lasso for d on Z to select a subset $Z^{\star,d} \in Z$ that best predicts d
   - 2.2 Run OLS for d on the selected $Z^{\star,d} \in Z \Longrightarrow$ Get residuals $\dot{d}$
3. Step 3:
   - 3.1 Run an OLS for $\dot{y}$ on $\dot{d}$

Partialling-out via ~~Post-Lasso~~ ANY ML?

WHY POST-LASSO?

The Partialling-out via Post-Lasso procedure in three steps:

1. Step 1:
   1.1 Run Lasso for ~~y on Z to select a subset $Z^\star$~~ ANY ML predicts y
   1.2 Run OLS for ~~y on the selected $Z^{\star,y}$~~ $\in Z \Longrightarrow$ Get residuals $\dot{y}$
2. Step 2:
   2.1 Run Lasso for ~~d on Z to select a subset $Z^\star$~~ ANY ML predicts d
   2.2 Run OLS for ~~d on the selected $Z^{\star,d}$~~ $\in Z \Longrightarrow$ Get residuals $\dot{d}$
3. Step 3:
   3.1 Run an OLS for $\dot{y}$ on $\dot{d}$

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
- $Z_i = (z_{1,i}, \ldots, z_{p,i})$ - vector of controls

Partially Linear Regression setting

Outcome Parameter of interest (ATE)

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

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

Partially Linear Regression setting

Outcome Parameter of interest (ATE)

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

Treatment equation: Because we assume that d_i is exogenous conditional on $g_0(Z_i)$

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

- y_i - outcome variable
- d_i - treatment variable
- $Z_i = (z_{1,i}, \dots, z_{p,i})$ - vector of controls

Partially Linear Regression setting

Outcome equation:

nuisance function $g(.)$

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

Treatment equation:

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

nuisance function $m(.)$

- y_i - outcome variable
- d_i - treatment variable
- $Z_i = (z_{1,i}, \dots, z_{p,i})$ - vector of controls

Partially Linear Regression setting

Outcome equation:

Captures the relationship between the nuisance parameters and y

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

Treatment equation:

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

Captures the relationship between the nuisance parameters and d

- y_i - outcome variable
- d_i - treatment variable
- $Z_i = (z_{1,i}, \dots, z_{p,i})$ - vector of controls

Double De-biased Machine Learning estimator

1. Use ML* to predict y from $z \Longrightarrow$ get $\hat{u} = y - \hat{g}_0(Z)$
2. Use ML* to predict d from $z \Longrightarrow$ 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, Random Forest, etc.

The estimator:

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

$$y_i = \beta_0 d_i + g_0(Z_i) + u_i,$$
$$d_i = m_0(Z_i) + v_i,$$

DML bias

$$E(\hat{\beta}_0 - \beta) = a + b + c$$

$$a \xrightarrow{p} 0 \text{ if } E(u_i v_i) = 0$$

Conditional Independence Assumption:
If treatment is as good as random after we control for nuisance parameters

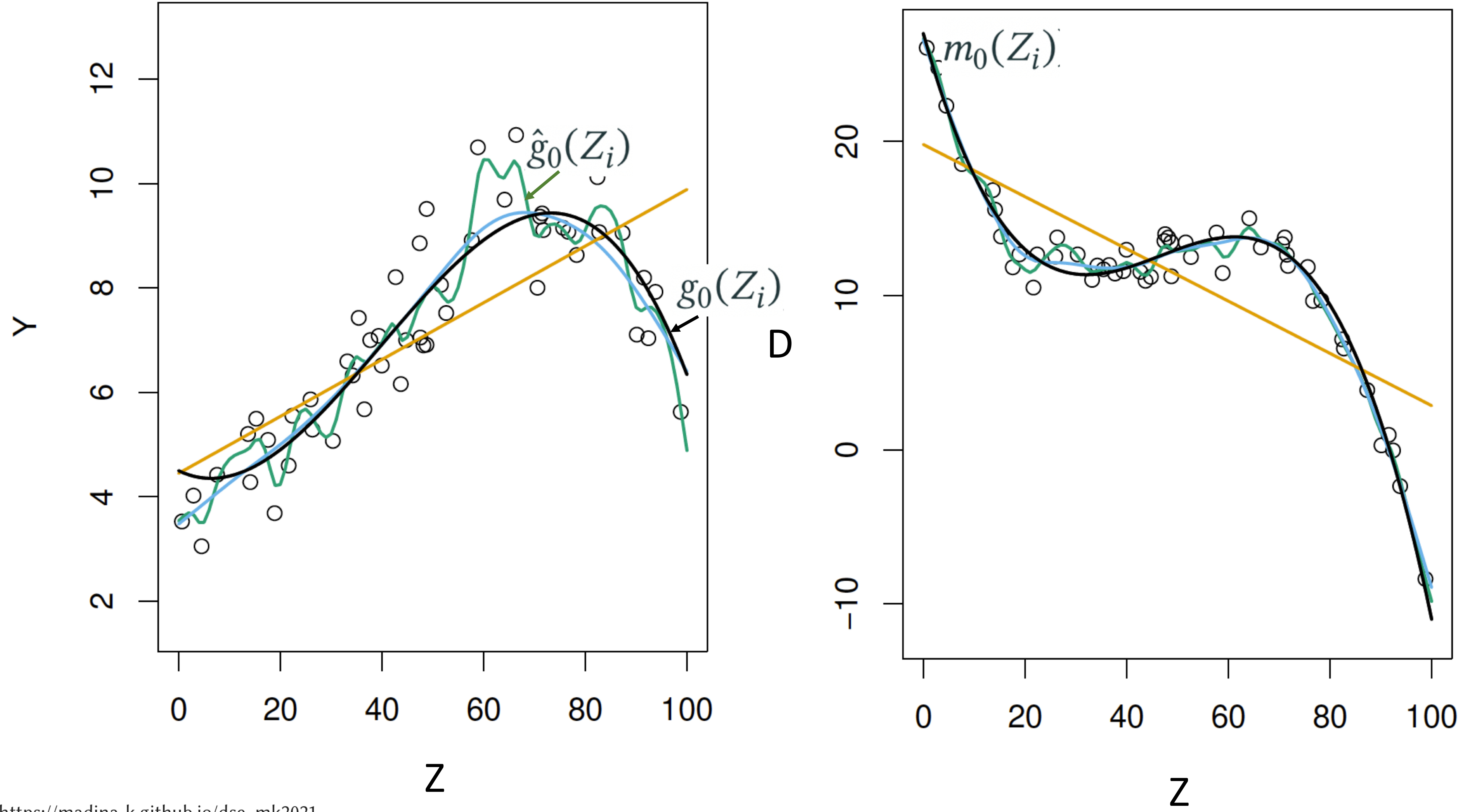
$$y_i = \beta_0 d_i + g_0(Z_i) + u_i,$$
$$d_i = m_0(Z_i) + v_i,$$

DML bias

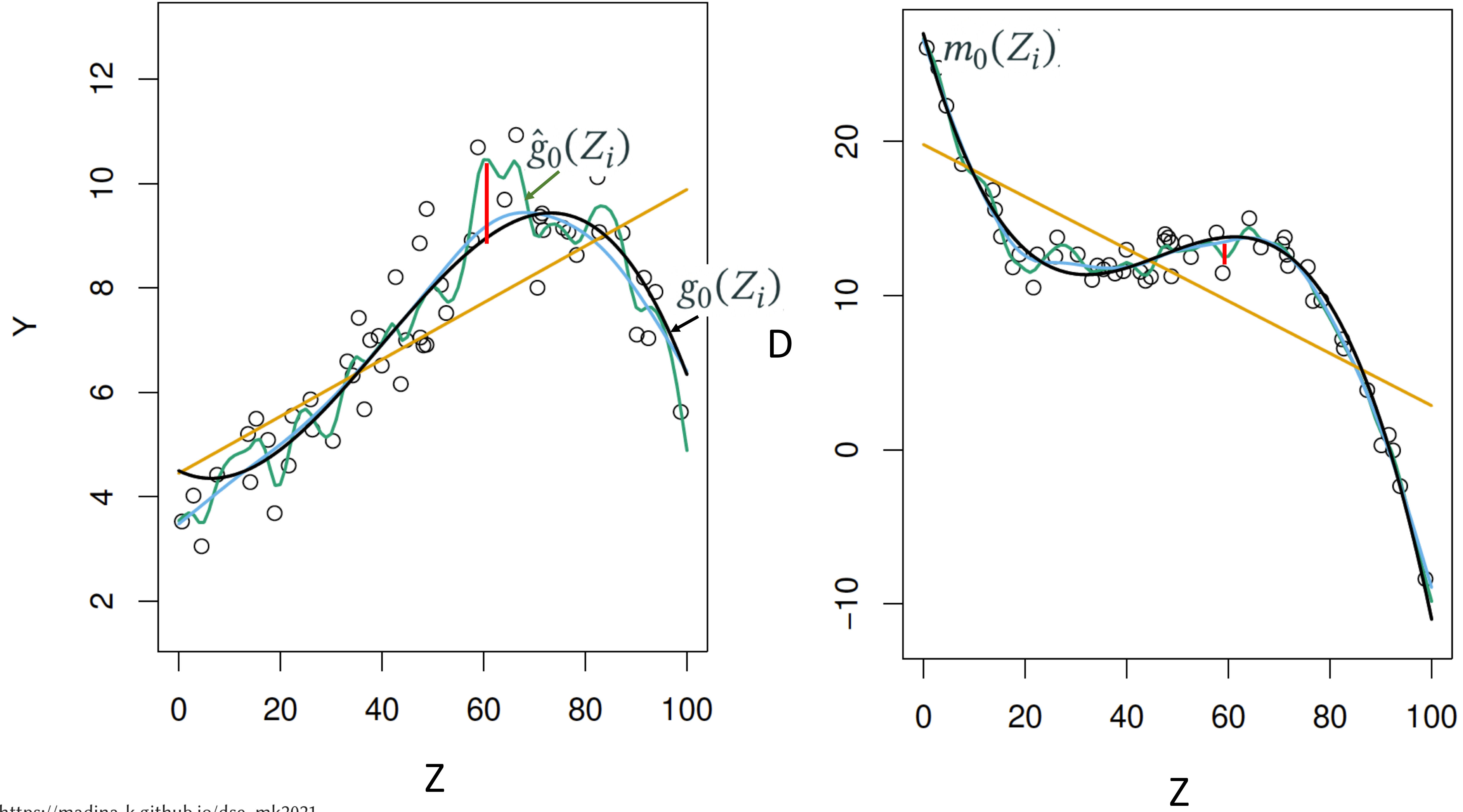
$$E(\hat{\beta}_0 - \beta) = a + b + c$$

$$a \xrightarrow{p} 0 \text{ if } E(u_i v_i) = 0$$

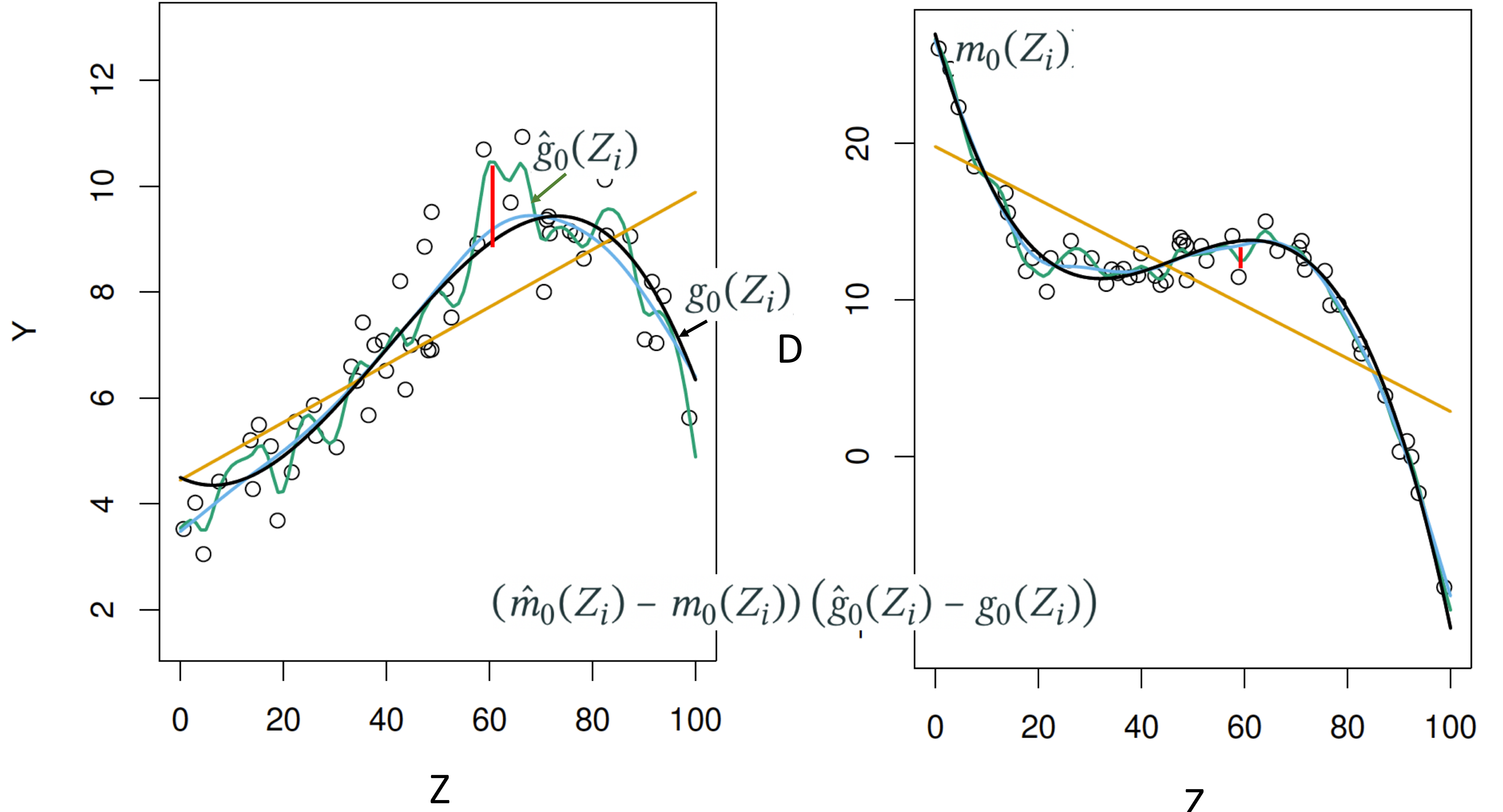
$$b \xrightarrow{p} 0 \text{ if } E\left((\hat{m}_0(Z_i) - m_0(Z_i))\left(\hat{g}_0(Z_i) - g_0(Z_i)\right)\right) = 0$$



$\hat{g}_0(Z_i)$
$g_0(Z_i)$
Y
Z
D
$m_0(Z_i)$
Z



https://madina-k.github.io/dse_mk2021



https://madina-k.github.io/dse_mk2021

DML bias

$$y_i = \beta_0 d_i + g_0(Z_i) + u_i,$$
$$d_i = m_0(Z_i) + v_i,$$

$$E(\hat{\beta}_0 - \beta) = a + b + c$$

$$a \xrightarrow{p} 0 \text{ if } E(u_i v_i) = 0$$

$$b \xrightarrow{p} 0 \text{ if } E\left((\hat{m}_0(Z_i) - m_0(Z_i))\left(\hat{g}_0(Z_i) - g_0(Z_i)\right)\right) = 0$$

$$c \xrightarrow{p} 0 \text{ if } E\left(v_i(\hat{g}_0(Z_i) - g_0(Z_i))\right) = 0$$

Bias induced by overfitting

$$y_i = \beta_0 d_i + g_0(Z_i) + u_i,$$
$$d_i = m_0(Z_i) + v_i,$$

DML bias

$$E(\hat{\beta}_0 - \beta) = a + b + c$$

$$a \xrightarrow{p} 0 \text{ if } E(u_i v_i) = 0$$

$$b \xrightarrow{p} 0 \text{ if } E\left((\hat{m}_0(Z_i) - m_0(Z_i))\left(\hat{g}_0(Z_i) - g_0(Z_i)\right)\right) = 0$$

$$c \xrightarrow{p} 0 \text{ if } E\left(v_i(\hat{g}_0(Z_i) - g_0(Z_i))\right) = 0$$

To guarantee it is zero, do not estimate β on the same sample on which you fit *g()* and *m()* functions

TRAINING DATA

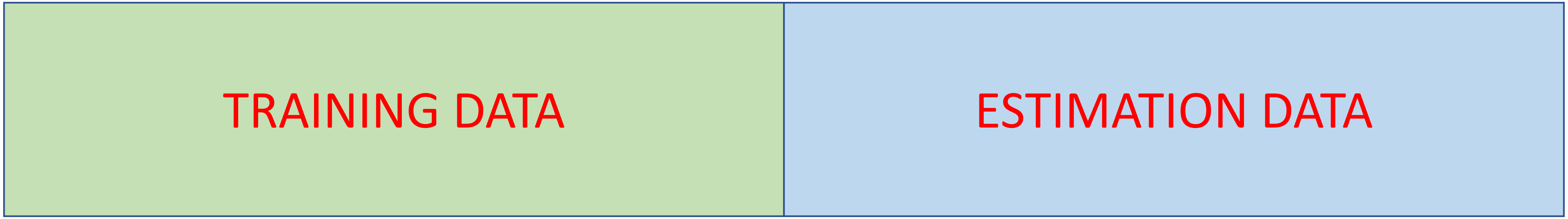
(used to fit *g(Z)* and *m(Z)*)

ESTIMATION DATA

(used to estimate final OLS)

$\hat{\beta}_0$ should be estimated using out-of-sample predictions

Cross-fitting DML

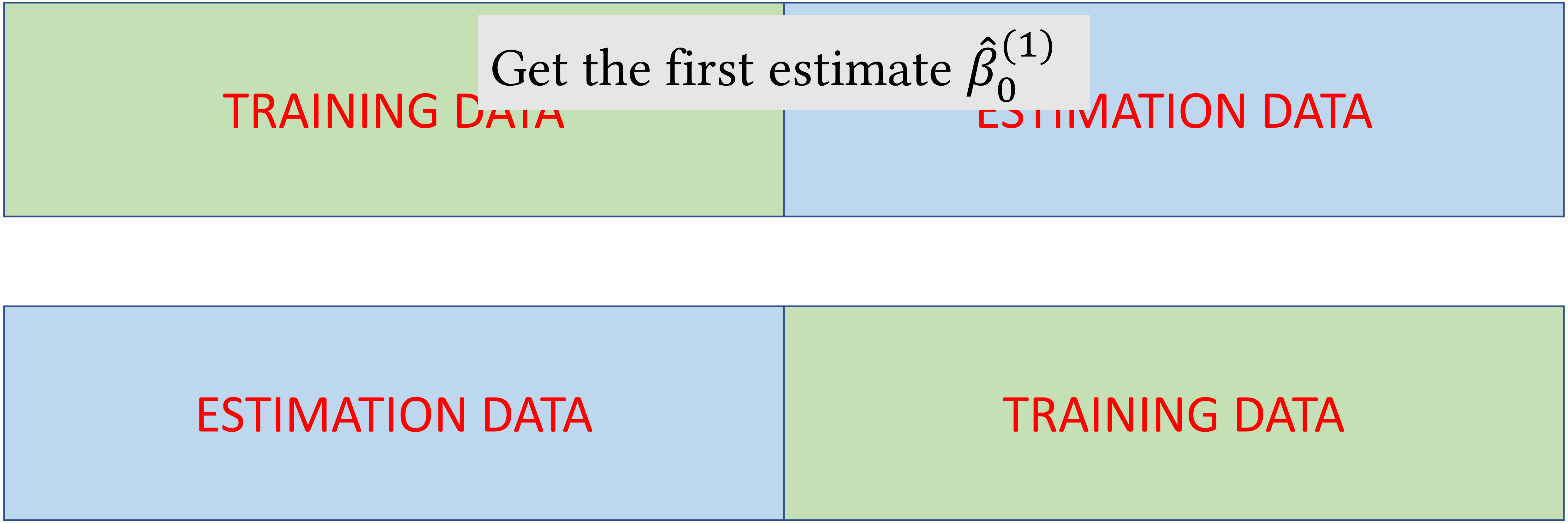


Cross-fitting DML

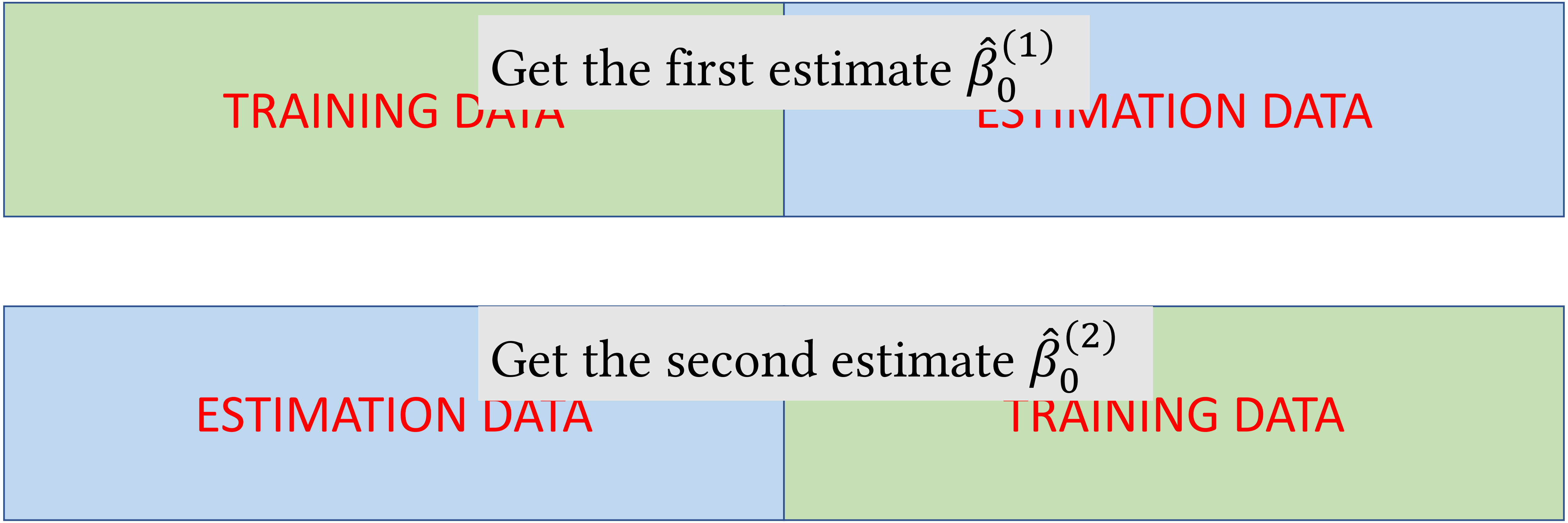
TRAINING DATA | ESTIMATION DATA

Get the first estimate $\hat{\beta}_0^{(1)}$

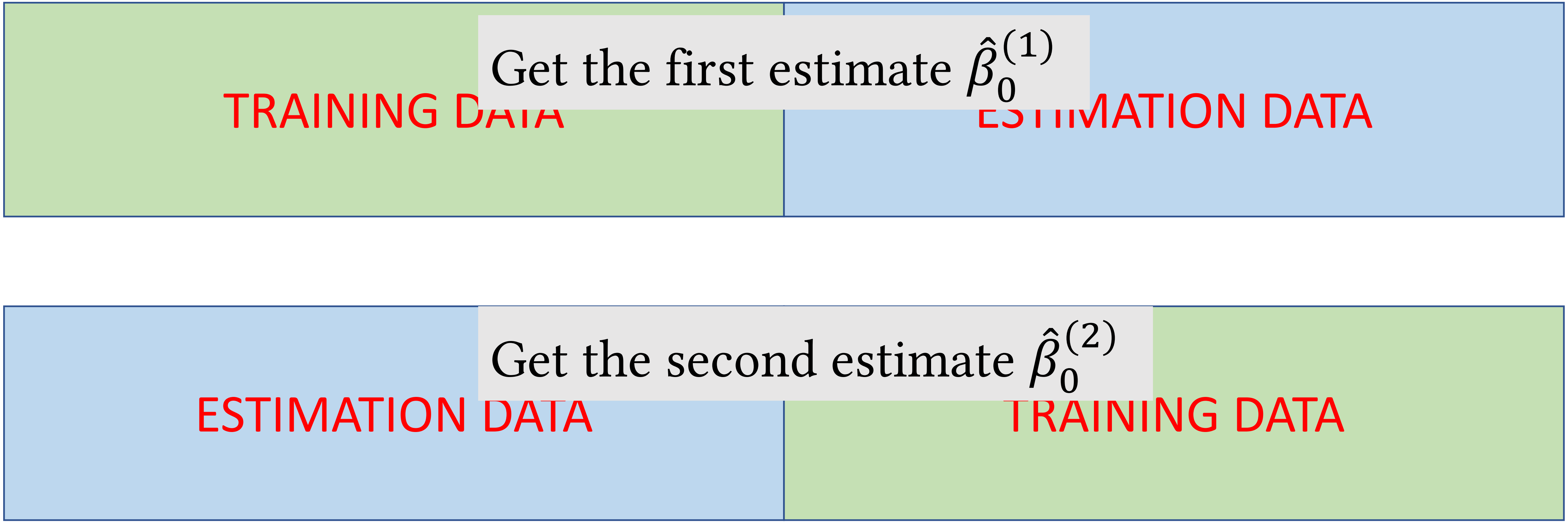
Cross-fitting DML



Cross-fitting DML



Cross-fitting DML



Cross-fitted DML estimator $\hat{\beta} = 0.5\left(\hat{\beta}_0^{(1)} + \hat{\beta}_0^{(2)}\right)$