Simulation Plan

Lingxiao Wang

11/28/2018

Simulation design in Setoguchi et al. 2008

Variables

- Binary exposure $A \sim Bernoulli(q)$ (i.e. group indicator, =1 for treatment group, =0 for control group)
- Outcome variable $y \sim Bernoulli(\mu)$
- 10 covariates $W_1 - W10$
  - $W_1 - W_4$: independently associated with both A and Y
  - $W_5 - W_7$: associated with A only
  - $W_8 - W_{10}$: associated with Y only

True propensity score models

1. Additive + linear model

$$logit(q) = \beta_0 + \beta_1 W_1 + \beta_2 W_2 + \beta_3 W_3 + \beta_4 W_4 + \beta_5 W_5 + \beta_6 W_6 + \beta_7 W_7$$

2. Additive + mild non-linearity model

$$logit(q) = \beta_0 + \beta_1 W_1 + \beta_2 (W_2 + W_2^2) + \beta_3 W_3 + \beta_4 W_4 + \beta_5 W_5 + \beta_6 W_6 + \beta_7 W_7$$

3. Additive + moderate non-linearity model

$$\begin{aligned} logit(q) =& \beta_0 + \beta_1 W_1 + \beta_2 (W_2 + W_2^2) + \beta_3 W_3 + \beta_4 (W_4 + W_4^2) + \beta_5 W_5 + \beta_6 W_6 + \\ & \beta_7 (W_7 + W_7^2) \end{aligned}$$

4. Mild non-additive + linearity model

$$\begin{aligned} logit(q) =& \beta_0 + \beta_1 W_1 + \beta_2 W_2 + \beta_3 W_3 + \beta_4 W_4 + \beta_5 W_5 + \beta_6 W_6 + \beta_7 W_7 + \\ & \beta_8 W_1 W_3 + \beta_9 W_2 W_4 + \beta_1 0 W_4 W_5 + \beta_1 1 W_5 W_6 \end{aligned}$$

5. Mild non-additive + mild non-linearity model

$$\begin{aligned} logit(q) =& \beta_0 + \beta_1 W_1 + \beta_2 (W_2 + W_2^2) + \beta_3 W_3 + \beta_4 W_4 + \beta_5 W_5 + \beta_6 W_6 + \beta_7 W_7 + \\ & \beta_8 W_1 W_3 + \beta_9 W_2 W_4 + \beta_{10} W_4 W_5 + \beta_{11} W_5 W_6 \end{aligned}$$

6. Moderate non-additive + moderate linearity model

$$\begin{aligned} logit(q) =& \beta_0 + \beta_1 W_1 + \beta_2 (W_2 + W_2^2) + \beta_3 W_3 + \beta_4 (W_4 + W_4^2) + \beta_5 W_5 + \beta_6 W_6 + \\ & \beta_7 (W_7 + W_7^2) + \beta_8 W_1 W_3 + \beta_9 W_2 W_4 + \beta_{10} W_3 W_5 + \beta_{11} W_4 W_6 + \beta_{12} W_5 W_7 + \\ & \beta_1 3 W_1 W_6 + \beta_1 4 W_2 W_3 + \beta_1 5 W_4 W_5 + \beta_1 6 W_5 W_6 \end{aligned}$$

Outcome model

$$logit(\mu) = \alpha_0 + \alpha_1 W_1 + \alpha_2 W_2 + \alpha_3 W_3 + \alpha_4 W_4 + \alpha_8 W_8 + \alpha_9 W_9 + \alpha_{10} W_{10} + A$$

Our simulation

Modifications

- PPS sampling by q and $1 - q$ for cohort and survey sample respectively
- Missing some important covariates in propensity score model
  Replace some of variable in confounders $W_1 - W_4$ using (highly) correlated variables $(W_1' - W_4')$

Applying methods

- Logistic regression (main effects only, main effects + two-way interactions)
- Machine learning methods