Comparative Effectiveness if Machine Learning Methods for Causal Inference in Agricultural Economics

Syed Fuad

Agricultural & Applied Economics Association Annual Meeting,

Washington DC: July 23-25, 2023

Introduction

- Machine learning:
  - ML is well-developed and widely used nonparametric prediction methods that work well with big data.
  - Focus on prediction and applications of prediction.
- This is all very good for business applications where you care about predictions and not the causal relationship.
- Many problems in social sciences entail causal inference.
  - BUT when we ask causal questions, we are usually interested in one parameter of interest – the treatment effect?
- Up until recently, existing ML approaches to estimation, model selection and robustness do not directly apply to the problem of estimating causal parameters.

New developments: causal ML models

- Bayesian additive regression trees (Chipman et al., 2010; Hill, 2011)
- Double ML (Chernozhukov et al., 2017)
- Causal Forest (Athey and Wager, 2019)
- Bayesian Causal Forest (Caron et al., 2020)
- PSM with underlying ML models
- Many more!

Analysis setup

- Estimation methods that combine ML approaches for prediction component of models with causal approaches
- Causal ML models to measure Average Treatment Effect (ATE):
  - Propensity Score Matching (PSM): Propensity score estimation is a pure prediction problem
  - Double Machine Learning (DML): The method reduces the problem to first estimating two predictive tasks: (i) Predicting the outcome from the controls; (ii) Predicting the treatment from the controls; and (iii) Linear regression on residuals: resid(Y) ~ resid(D)
  - Causal Forest (CF): Extension of Random Forests. In random forests, the data is repeatedly split in order to minimize prediction error of an outcome variable. In causal forests instead of minimizing prediction error, data is split in order to maximize the difference across splits in the relationship between an outcome variable and a “treatment” variable
- Heterogenous Treatment Effect (HTE): X-learner

Data generating process

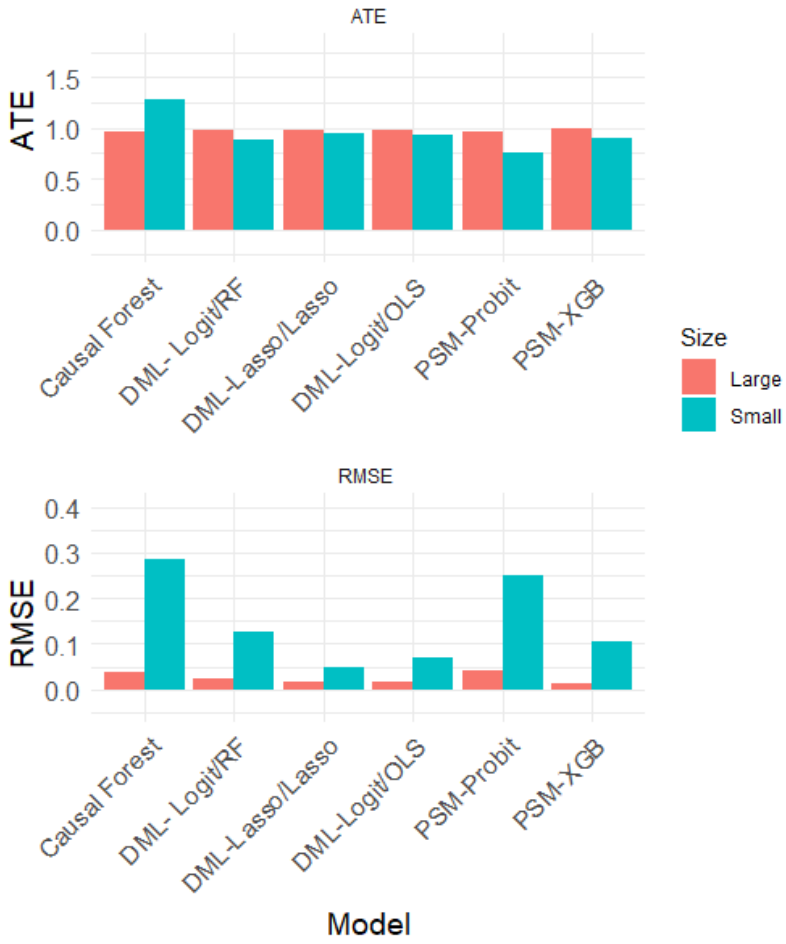
$$y_i = \theta d_i + x_i'\beta + u_i$$
$$d_i = x_i'\beta + v_i$$

- d_i represents the binary treatment variable (approximately 50% of the observations receive treatment)
- x_i represents a vector of k covariates, generated from a multivariate normal distribution; β is a vector of k parameters
- Increase sample size ($n = 150, 500, 5000, 15000$)
- Increase number of covariates ($k = 10, 50, 100$)
- Impose treatment heterogeneity ($\theta = 1;\ \theta \sim Normal(1,1)$)
- Change structure of data ($y_i = \theta d_i + x_i'\beta + u_i \ and \ y_i = \theta d_i + sin(x_i'\beta) + u_i$)

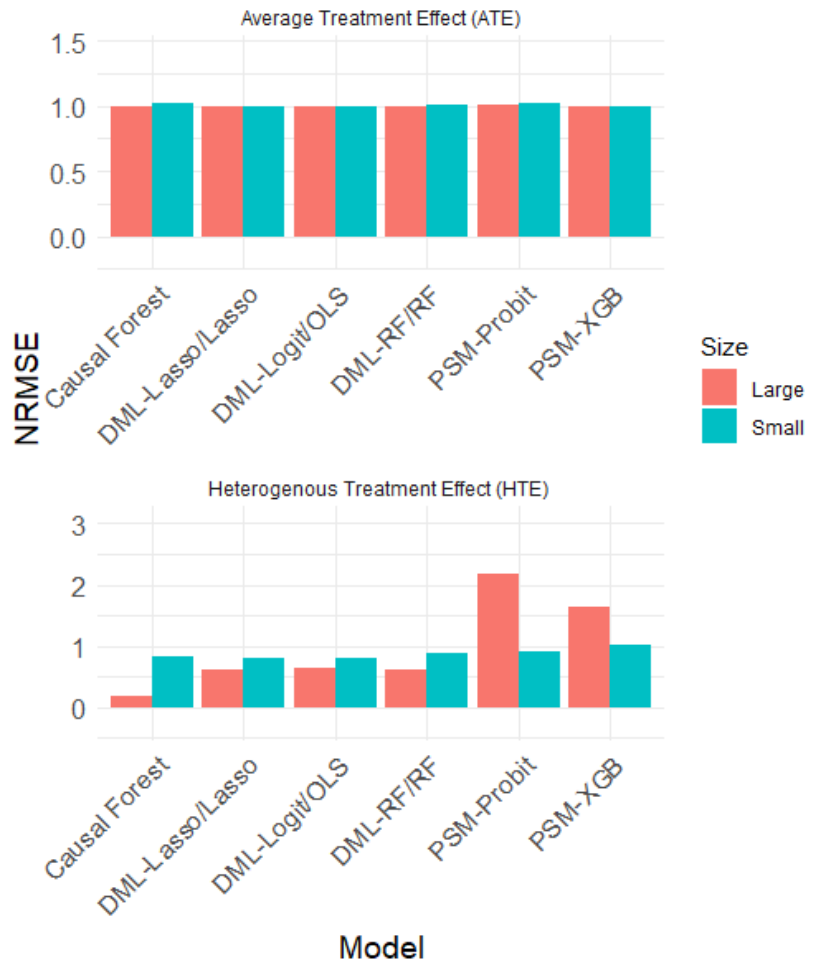
Results

$n = 150; k = 10;\ \theta = 1;\ y_i = \theta d_i + x_i'\beta + u_i$

$n = 15{,}000; k = 10;\ \theta = 1;\ y_i = \theta d_i + x_i'\beta + u_i$

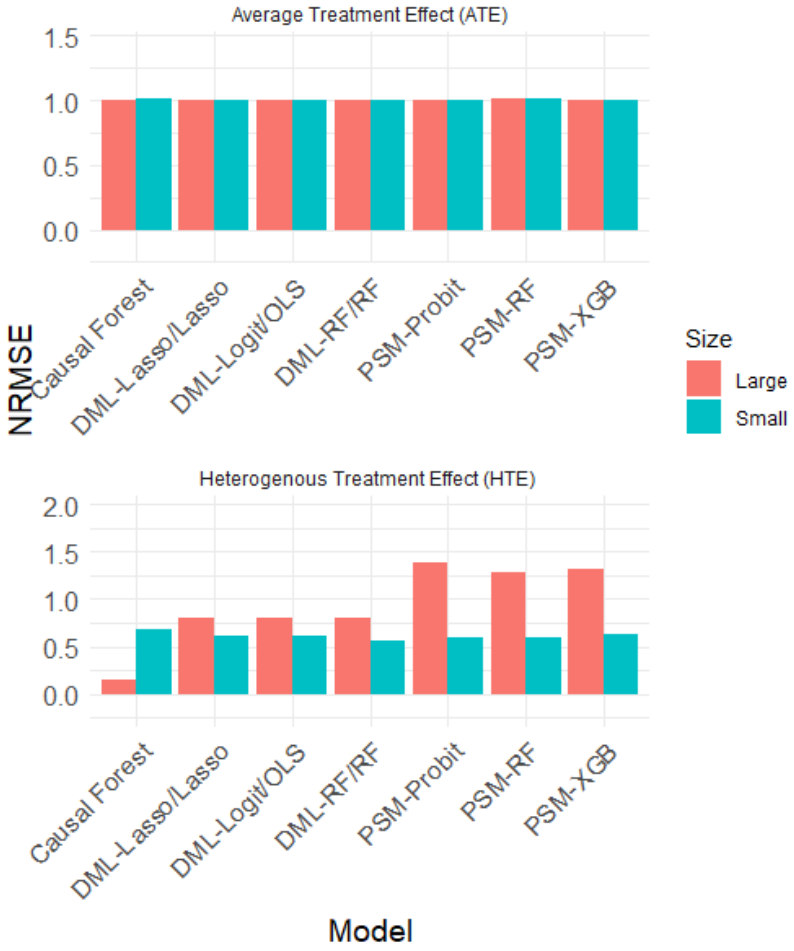


$n = 15{,}000; k = 10; \theta \sim Normal(1,1);\ y_i = \theta d_i + x_i'\beta + u_i$

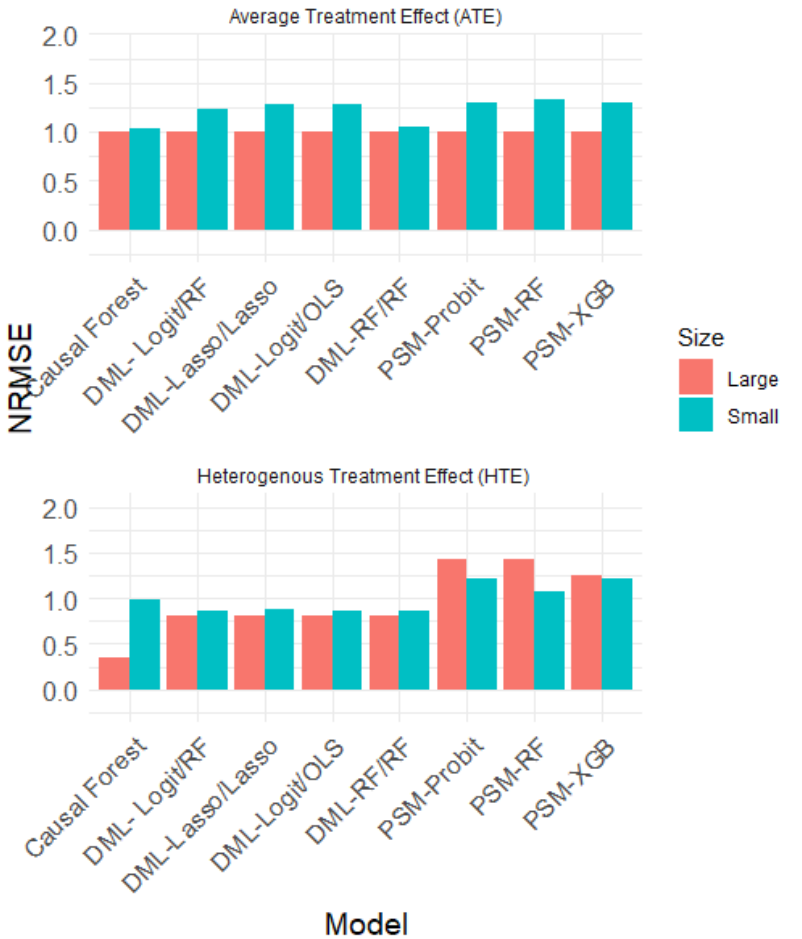


Results (contd.)

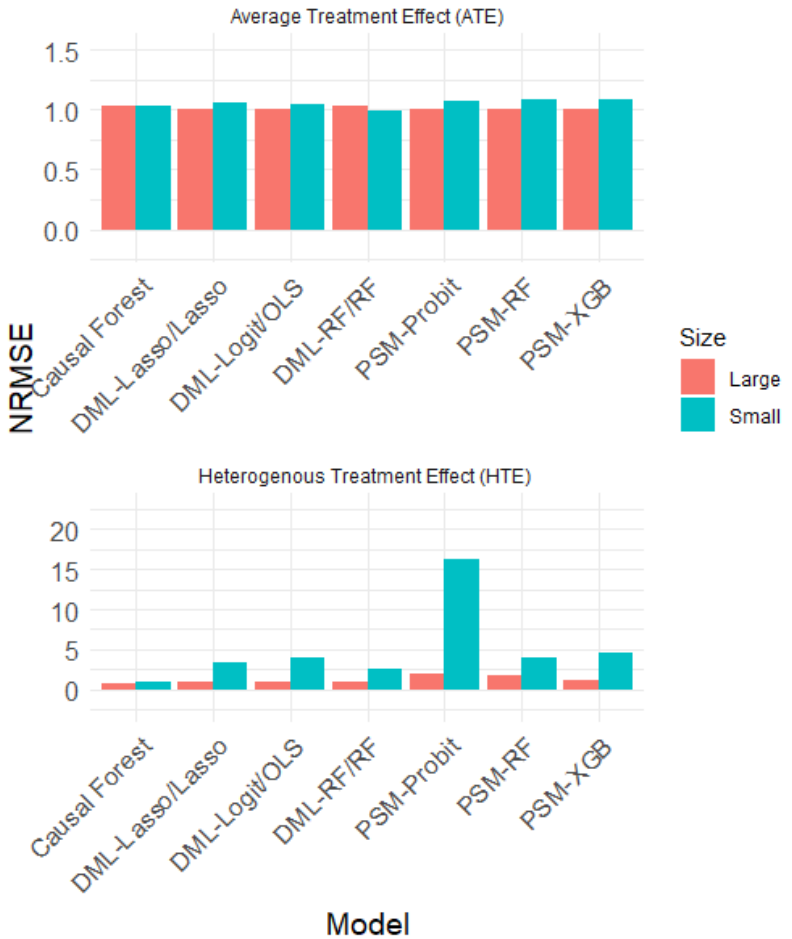
$n = 15{,}000; k = 10; \theta \sim Normal(1,1);$
$y_i = \theta d_i + sin(x_i'\beta) + u_i$



$n = 15{,}000; k = 100; \theta \sim Normal(1,1);$
$y_i = \theta d_i + sin(x_i'\beta) + u_i$



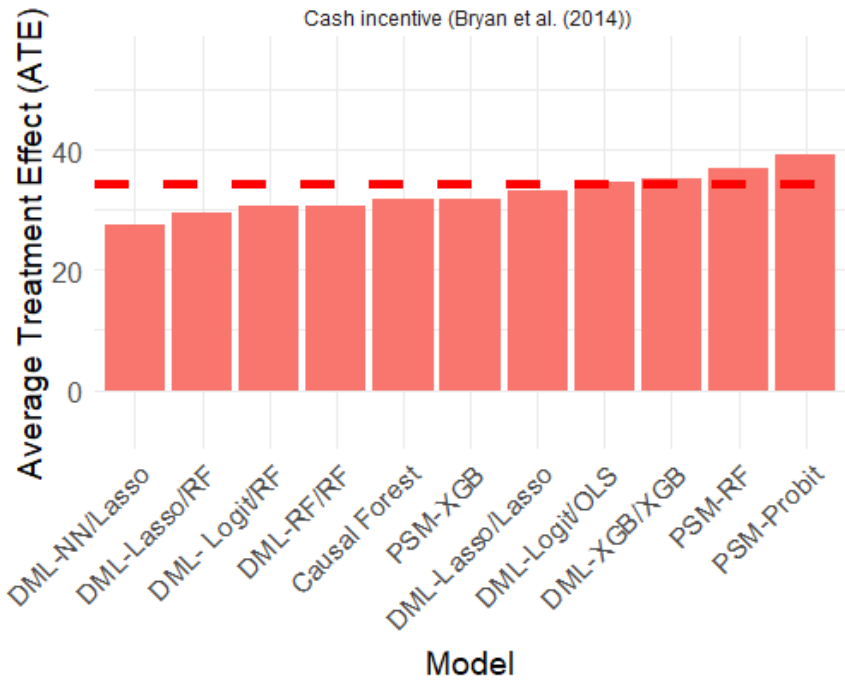
$n = 15{,}000; k = 100; \theta \sim Normal(1,1);\ y_i = \theta d_i + x_i'\beta + u_i$



Results (contd.)

-0.53 ± 1.16

-0.71 ± 1.16



34.27
(23.08)

Conclusion

- No one-size-fits-all model
- Complex models are not necessarily always effective
  - May be well worth the effort to not only experiment with a range of hyperparameters, but also with different underlying ML model frameworks
- When measuring ATE, DML is generally advantageous over both PSM-ML and CF, and this pattern remains consistent even as data size increases, although the relative outperformance of DML over the other methods decline
- Among the best performers are DML methods that use simple underlying algorithms, such as DML-Lasso/Lasso, and DML-Logit/OLS
- If interested in estimating HTE, DML marginally outperforms PSM and CF, especially at small sample sizes
  - As sample size increases, CF shows superior performance over both DML and PSM