Observational studies with multi-level treatments

By

Indrajith Wasala Mudiyanselage

11/28/2022

Outline:

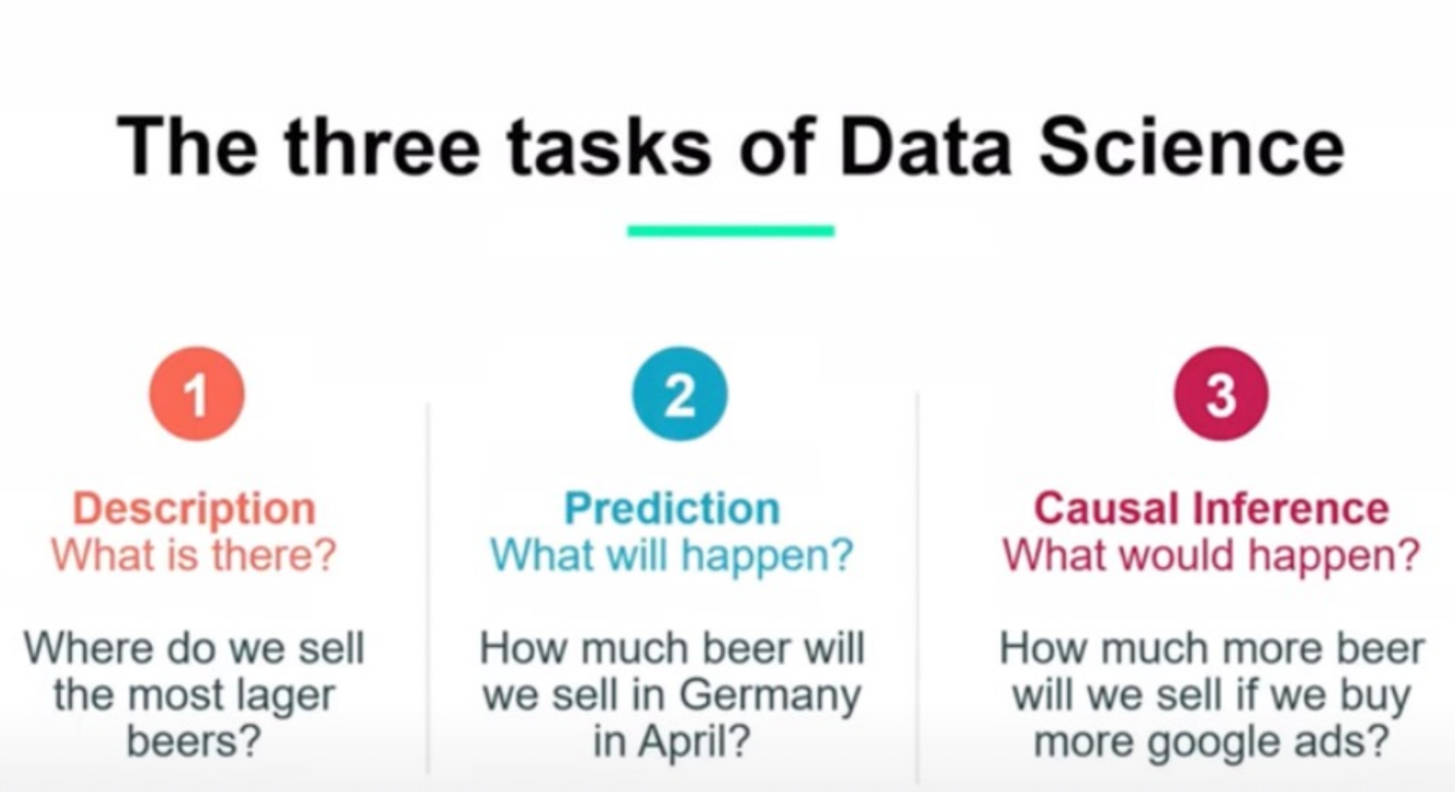
- Introduction & Motivation
- Existing literature
- Simulation setting & results
- A real example
- Discussion

Introduction & Motivation

Introduction:

The goal of causal inference is to answer certain questions based on the causal structure of the problem.

Example



The three tasks of Data Science

1

Description
What is there?

Where do we sell the most lager beers?

2

Prediction
What will happen?

How much beer will we sell in Germany in April?

3

Causal Inference
What would happen?

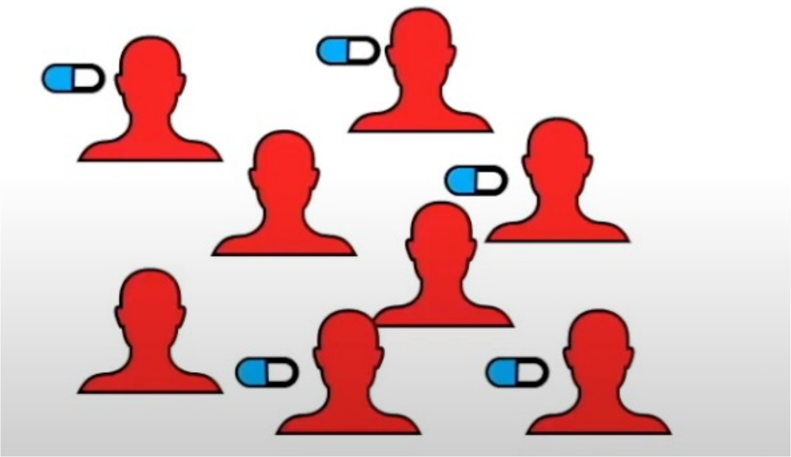
How much more beer will we sell if we buy more google ads?

To get an answer we need to estimate the causal effect of google ads on beer sales.

In an investigation of the causal effect of a treatment, one of the most common issues is that the groups of subjects in the study are not randomly assigned. Then there is no interference or manipulation of the research subjects and no control over treatment groups. These studies are called observational studies.

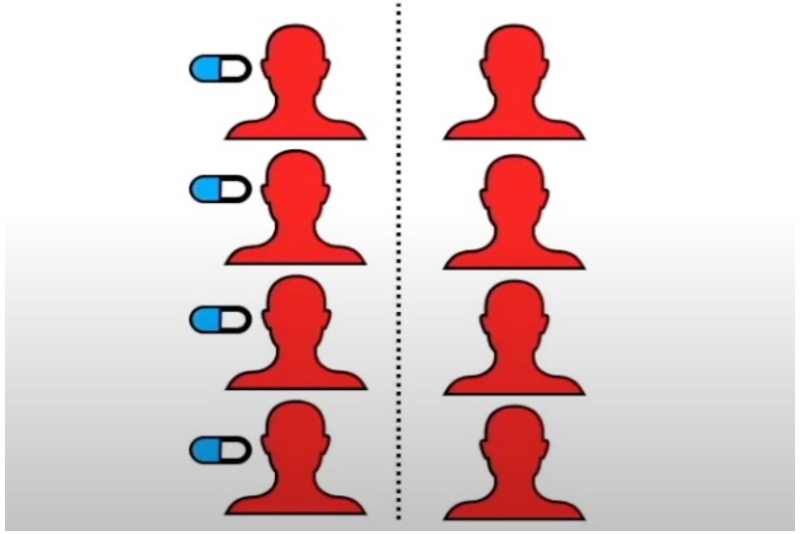
Observational Studies

Passively measuring without any intervention in data generating process



Experimental studies

Intentional *manipulation* of data generating process for a particular goal



Confounders:

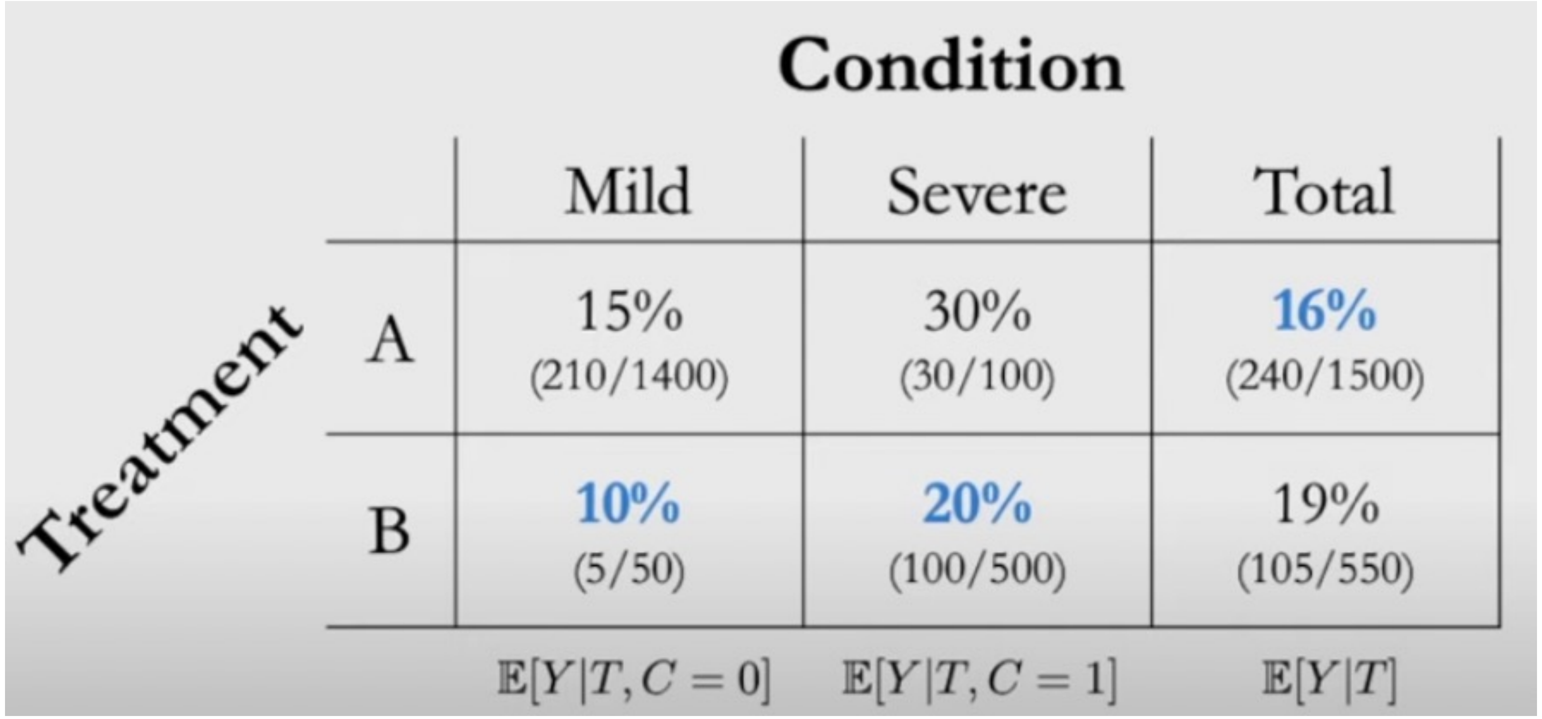
- Confounders are the variables associated with both the treatment and the outcome.
- When Confounders are present, non-random group assignments could result in biased estimates of the treatment effect.
- Example

Consider a new disease (Say COVID-50)

Treatment T: A (0) and B (1)

Condition C: Mild (0) or Severe (1)

Outcome Y: Alive (0) or Dead (1)



		Condition		
		Mild	Severe	Total
Treatment	A	15% (210/1400)	30% (30/100)	16% (240/1500)
	B	10% (5/50)	20% (100/500)	19% (105/550)
		$\mathbb{E}[Y \mid T, C=0]$	$\mathbb{E}[Y \mid T, C=1]$	$\mathbb{E}[Y \mid T]$

- There are many popular literatures to overcome this problem.

- Among the literature, Propensity score-based methods have been widely improved to adjust for confounders in observational studies to estimate causal treatment effects.

- The initial work on propensity scores focused on the case of binary treatments.

- Recent work generalizes the Propensity score-based methods to treatments with more than two levels (Multi-level treatment).

Existing literature

Causal inference for multi-level treatments with machine-learned propensity scores

- In this paper, the authors generalize Propensity score-based methods that have been developed to adjust for confounders in observational studies with binary treatments.

- In other words, improving existing binary treatment-based Propensity scores to the multi-level treatment case.

- Here instead of assuming an underlying binomial distribution for the treatment conditional on the covariates as in logistic regression, authors assume an underlying multinomial distribution.

Multinomial logistic regression (MLR)

Assume an underlying multinomial distribution then the generalized propensity score for each treatment level follows:

$$p(t|x)_{MLR} = \frac{1}{1 + \sum_{s=2}^{M} e^{\beta_s' x}} \text{ for } t = 1,$$

$$p(t|x)_{MLR} = \frac{e^{\beta_t' x}}{1 + \sum_{s=2}^{M} e^{\beta_s' x}} \text{ for } t = 2, \dots, M,$$

where $\beta_s = (1, \beta_{s1}, \dots, \beta_{sp})$ for $s = 2, \dots, M$ and p is the number of covariates.

Generalized boosted models (GBM)

- GBM is built on an ensemble of regression trees and each regression tree iteratively fits the residuals from the previous tree to approximate the propensity score function.

- The algorithm automatically includes interaction and nonlinear terms as the regression tree allows for multi-level splits.

- It is suggested we should fit a GBM that balance the covariates between the treatment t group and the entire sample as in the binary case.

Random forest (RF)

- RF is the aggregation of a collection of regression or classification trees fitted on bootstrap samples of the original dataset with the original sample size.
- A key feature of RF lies in that each tree in the tree ensemble is built on a random subset of the original covariates on a bootstrap sample of the original dataset to avoid overfitting.
- Given a vector of covariates, each tree votes for one class label. Then the generalized propensity score for treatment t can be estimated as the fraction of trees that classify/predict the given subject into treatment group t out of the entire collection of trees.
- If none of the trees predict a particular treatment level, then the RF estimator of the propensity score for that treatment will be zero.
- Unlike GBM in estimating the generalized propensity scores, RF uses a collection of classification trees instead of regression trees.

The data-adaptive matching score (DAMS)

- This is a weighted average of the propensity score estimates from a parametric model (such as a MLR model) and a nonparametric model (such as a RF model)
- This DAMS estimator can be denoted as,

$$\hat{p}(t|x)_{DAMS} = \lambda \hat{p}(t|x)_{MLR} + (1-\lambda)\hat{p}(t|x)_{RF},$$

$$\text{where } \lambda = \frac{\hat{p}(t|x)_{MLR}{}^{A(t)}[1-\hat{p}(t|x)_{MLR}]^{1-A(t)}}{\hat{p}(t|x)_{MLR}{}^{A(t)}[1-\hat{p}(t|x)_{MLR}]^{1-A(t)} + \hat{p}(t|x)_{RF}{}^{A(t)}[1-\hat{p}(t|x)_{RF}]^{1-A(t)}}.$$

- The bias and the variance of the treatment effect estimates can always be reduced, compared to MLR or RF models alone.

Treatment effect estimation

- The authors are interested in two types of causal effects: the average treatment effect (ATE) and the average treatment effect among the treated (ATT).
- The ATE of treatment t relative to treatment s is the difference of mean outcomes had the entire population been observed under t versus had the entire population been observed under s.
- The ATT would be the average treatment effect of treatment s among those treated with t is the difference in mean outcome among subjects who were treated with t versus the mean outcome they would have had if they received s.

$$\begin{aligned} ATE_{ts} &:= E[Y(t)] - E[Y(s)] \\ &= \mu_t - \mu_s . \end{aligned}$$

$$\begin{aligned} ATT_{t,ts} &:= E[Y(t)|A = t] - E[Y(s)|A = t] \\ &= \mu_{t,t} - \mu_{t,s} . \end{aligned}$$

There are various propensity score-based methods to estimate an ATE or an ATT including matching, stratification, inverse probability of treatment weighting (IPTW), etc. Here we focus on the IPTW method.

- Estimating ATEs

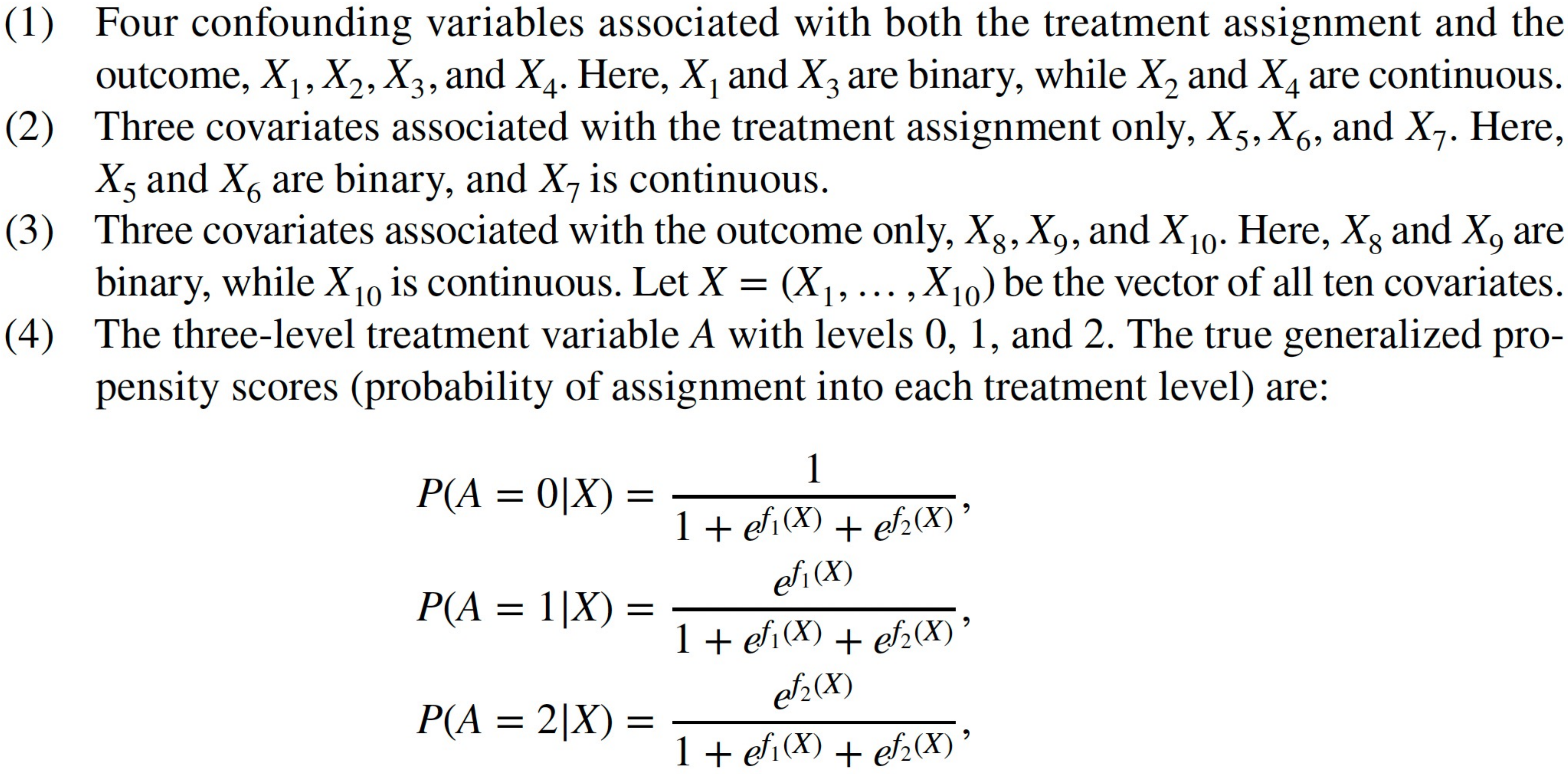
$$\widehat{ATE}_{ts} = \hat{\mu}_t - \hat{\mu}_s. \qquad \text{Where} \qquad \hat{\mu}_t = \frac{\sum_{i=1}^n A_i(t)w_i(t)Y_i}{\sum_{i=1}^n A_i(t)w_i(t)}. \quad w_i(t) = \frac{1}{\hat{p}_i(t|x)}.$$

- Estimating ATEs

$$\widehat{ATT}_{t,ts} = \hat{\mu}_{t,t} - \hat{\mu}_{t,s}. \qquad \text{Where} \qquad \hat{\mu}_{t,s} = \frac{\sum_{i=1}^n A_i(s)w_i(t,s)Y_i}{\sum_{i=1}^n A_i(s)w_i(t,s)}. \quad w_i(t,s) = \frac{\hat{p}_i(t|x)}{\hat{p}_i(s|x)}.$$

Simulation setting & results

For each simulated dataset, we have the following variables



(1) Four confounding variables associated with both the treatment assignment and the outcome, X_1, X_2, X_3, and X_4. Here, X_1 and X_3 are binary, while X_2 and X_4 are continuous.

(2) Three covariates associated with the treatment assignment only, X_5, X_6, and X_7. Here, X_5 and X_6 are binary, and X_7 is continuous.

(3) Three covariates associated with the outcome only, X_8, X_9, and X_{10}. Here, X_8 and X_9 are binary, while X_{10} is continuous. Let $X = (X_1, \ldots, X_{10})$ be the vector of all ten covariates.

(4) The three-level treatment variable A with levels 0, 1, and 2. The true generalized propensity scores (probability of assignment into each treatment level) are:

$$P(A = 0|X) = \frac{1}{1 + e^{f_1(X)} + e^{f_2(X)}},$$

$$P(A = 1|X) = \frac{e^{f_1(X)}}{1 + e^{f_1(X)} + e^{f_2(X)}},$$

$$P(A = 2|X) = \frac{e^{f_2(X)}}{1 + e^{f_1(X)} + e^{f_2(X)}},$$

(5) The continuous outcome variable Y with

$$E[Y|X] = -3.85 + 0.3X_1 - 0.36X_2 - 0.73X_3 - 0.2X_4 + 0.71X_8 - 0.19X_9 + 0.26X_{10} \\ - 0.4I(A = 1) - 0.7I(A = 2)$$

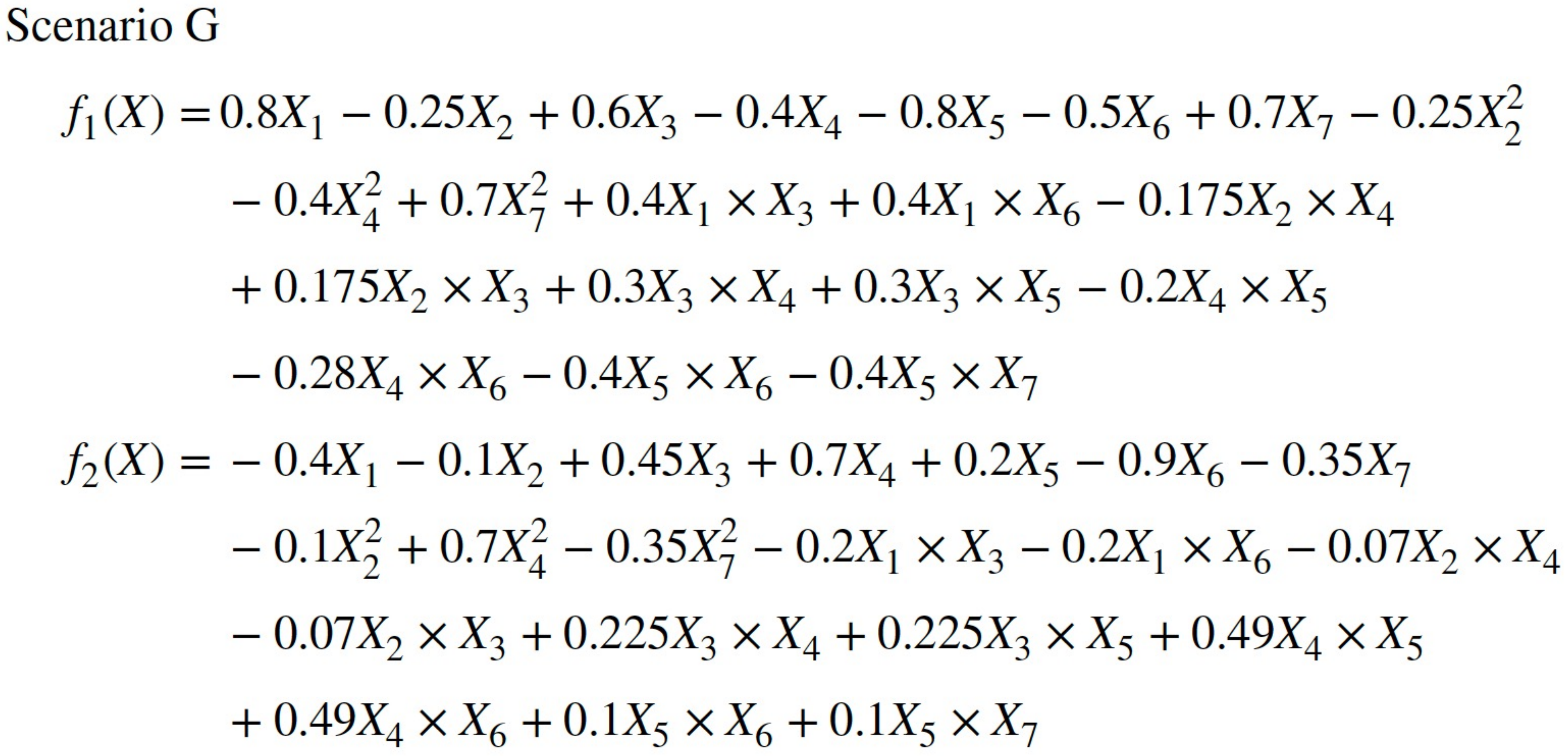
so we have the following true ATEs: $ATE_{10} = -0.4$, $ATE_{20} = -0.7$, and $ATE_{21} = -0.3$.

The covariates X_1, X_2, X_3, X_4, X_7, and X_{10} are generated through independent normal distributions with mean 0 and standard deviation 1. The rest of the covariates, X_5, X_6, X_8, X_9 are generated from normal distributions and have a correlation structure as follows:

$$corr(X_1, X_5) = 0.2,\ corr(X_2, X_6) = 0.9,\ corr(X_3, X_8) = 0.2,\ corr(X_4, X_9) = 0.9.$$

Then, X_1, X_3, X_5, X_6, X_8 and X_9 are dichtomized at the sample average (1 if the observation is greater than the sample average, 0 otherwise).

- Consider several scenarios (scenario A to scenario G) where the treatment assignment is related to the covariates with various degrees of non-linearity.
- The complexity of the relationships changes from simple scenario to more complex scenarios.



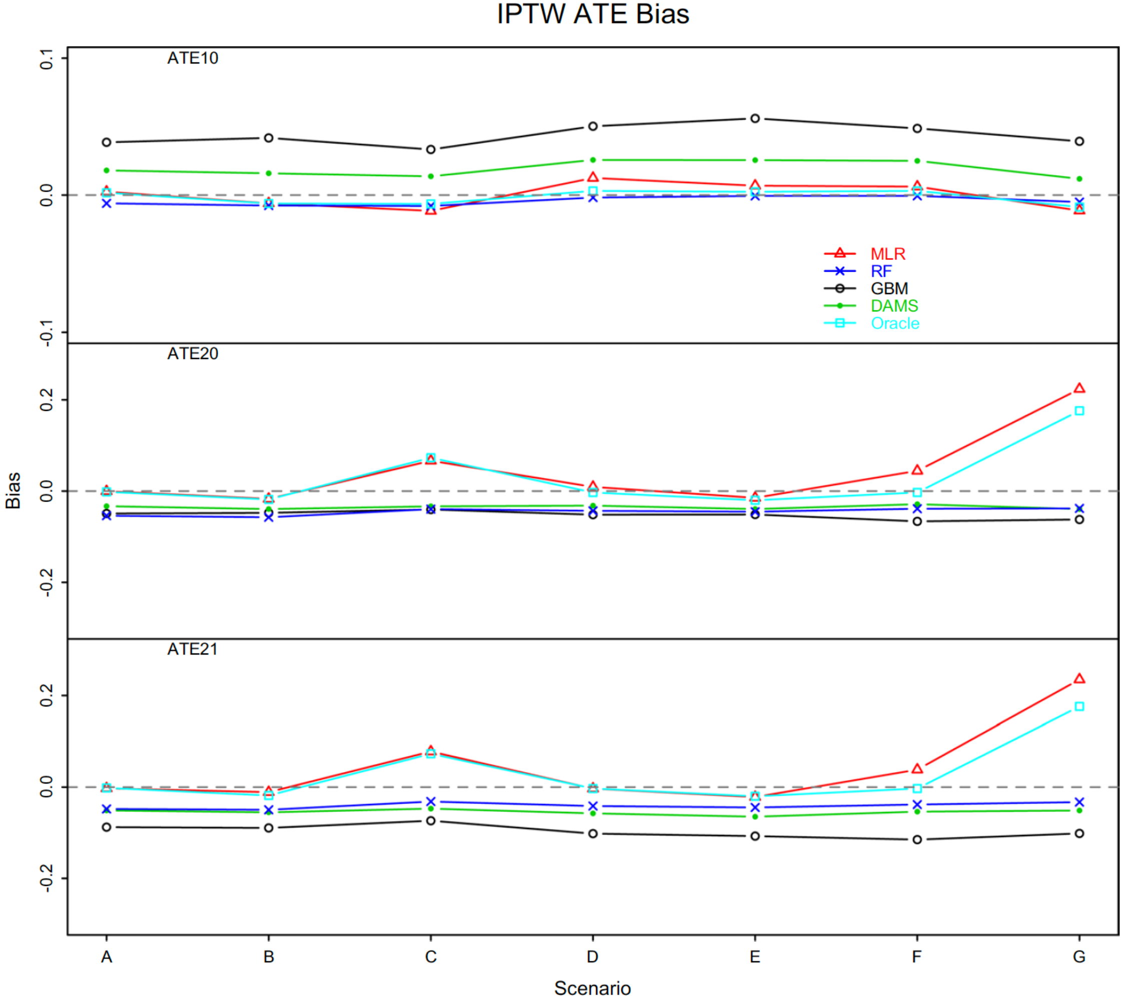
Scenario G

$$
\begin{aligned}
f_1(X) = {} & 0.8X_1 - 0.25X_2 + 0.6X_3 - 0.4X_4 - 0.8X_5 - 0.5X_6 + 0.7X_7 - 0.25X_2^2 \\
& - 0.4X_4^2 + 0.7X_7^2 + 0.4X_1 \times X_3 + 0.4X_1 \times X_6 - 0.175X_2 \times X_4 \\
& + 0.175X_2 \times X_3 + 0.3X_3 \times X_4 + 0.3X_3 \times X_5 - 0.2X_4 \times X_5 \\
& - 0.28X_4 \times X_6 - 0.4X_5 \times X_6 - 0.4X_5 \times X_7 \\
f_2(X) = {} & -0.4X_1 - 0.1X_2 + 0.45X_3 + 0.7X_4 + 0.2X_5 - 0.9X_6 - 0.35X_7 \\
& - 0.1X_2^2 + 0.7X_4^2 - 0.35X_7^2 - 0.2X_1 \times X_3 - 0.2X_1 \times X_6 - 0.07X_2 \times X_4 \\
& - 0.07X_2 \times X_3 + 0.225X_3 \times X_4 + 0.225X_3 \times X_5 + 0.49X_4 \times X_5 \\
& + 0.49X_4 \times X_6 + 0.1X_5 \times X_6 + 0.1X_5 \times X_7
\end{aligned}
$$

Simulation comparison (ATE)

- First, consider the bias, which is the difference between the average of the ATE estimates from the 1000 datasets and the respective true ATE.

$$Bias = \overline{\widehat{ATE}} - ATE$$



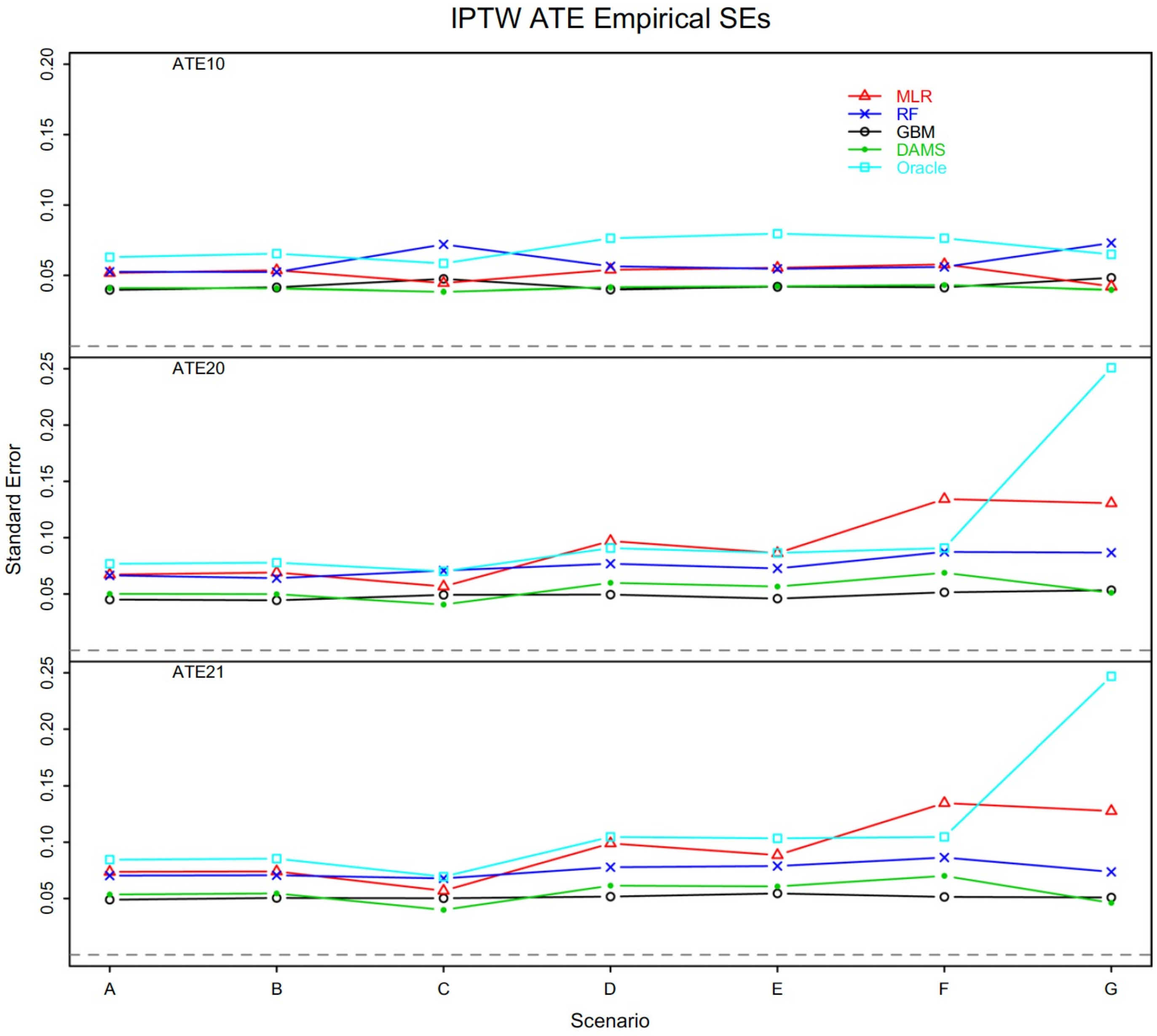
- Estimating generalized propensity scores by GBM seems to result in the most biased $\widehat{\text{ATE}}_{10,\text{IPTW}}$ in all seven scenarios, and this method seems to consistently overestimate the ATE.
- MLR and RF seem to result in lower bias in $\widehat{\text{ATE}}_{10,\text{IPTW}}$.
- The bias of $\widehat{\text{ATE}}_{20,\text{IPTW}}$ and $\widehat{\text{ATE}}_{21,\text{IPTW}}$ seem to be similar for each propensity score estimation method across all scenarios.
- MLR seems to result in a higher bias whereas the bias of the other three methods remain consistent across the different scenarios.
- When there is considerable non-linearity and non-additivity (like scenario G), the simple MLR model did not do an adequate job in estimating the propensity scores while machine learning methods perform well.

$$empirical\ SE = \sqrt{\frac{1}{999}\sum_{j=1}^{1000}(\widehat{ATE}_j - \overline{\widehat{ATE}})^2}.$$

$$averageSE = \frac{1}{1000}\sum_{j=1}^{1000}\widehat{SE}_j,$$

Method	Empirical standard error	Average standard error
MLR	0.052	0.078
RF	0.053	0.076
GBM	0.040	0.068
DAMS	0.041	0.071

Table 1 Comparison of empirical standard error and average standard error of $\widehat{ATE}_{10,IPTW}$ in scenario A

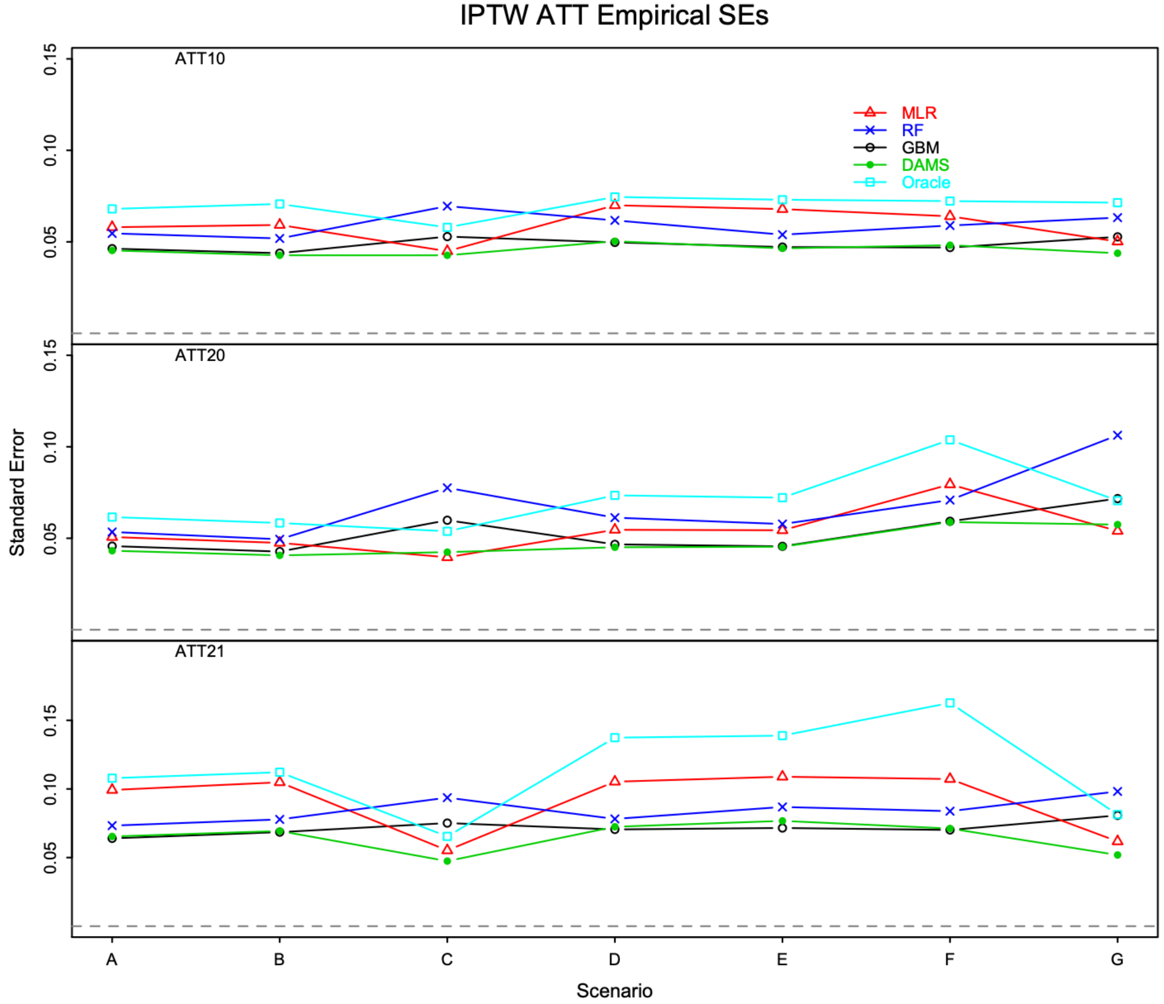


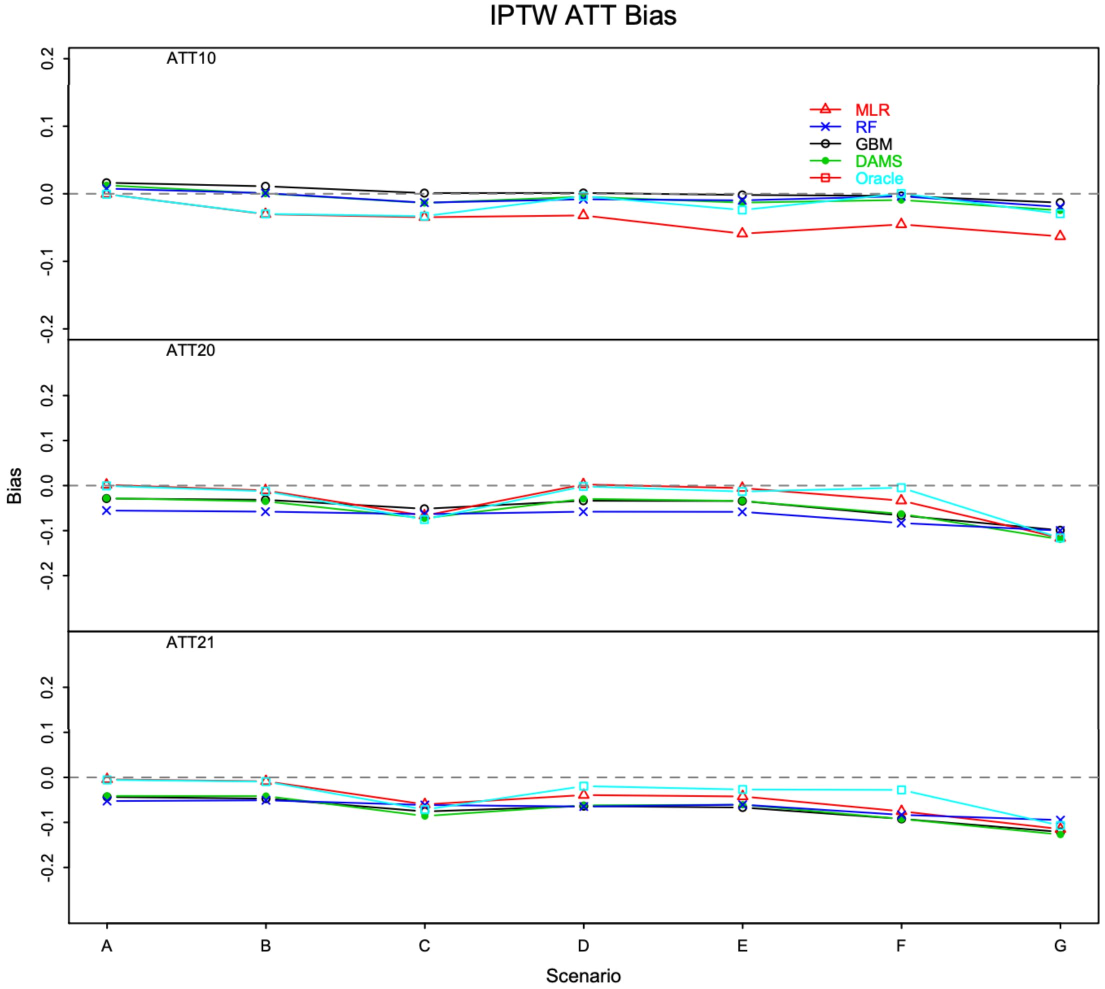
- The empirical SEs for $\widehat{\mathrm{ATE}}_{10,\mathrm{IPTW}}$ are similar between all methods and generally are constant across all scenarios. In scenarios C and G with non-linearity, RF seems to yield slightly higher empirical SEs.

- As was the case for bias, the empirical SE performance patterns for $\widehat{\mathrm{ATE}}_{20,\mathrm{IPTW}}$ and $\widehat{\mathrm{ATE}}_{21,\mathrm{IPTW}}$ are very similar. As the complexity of the scenarios increases, MLR yields higher empirical SEs compared to the other three methods.

- GBM and DAMS tend to result in similar levels of empirical SE that are lower compared to the other two methods.

Simulation comparison (ATT)


IPTW ATT Empirical SEs


IPTW ATT Bias

- The nonparametric algorithms, RF, DAMS and GBM lead to less biased estimates compared to MLR, when the MLR model is misspecified.

A real example

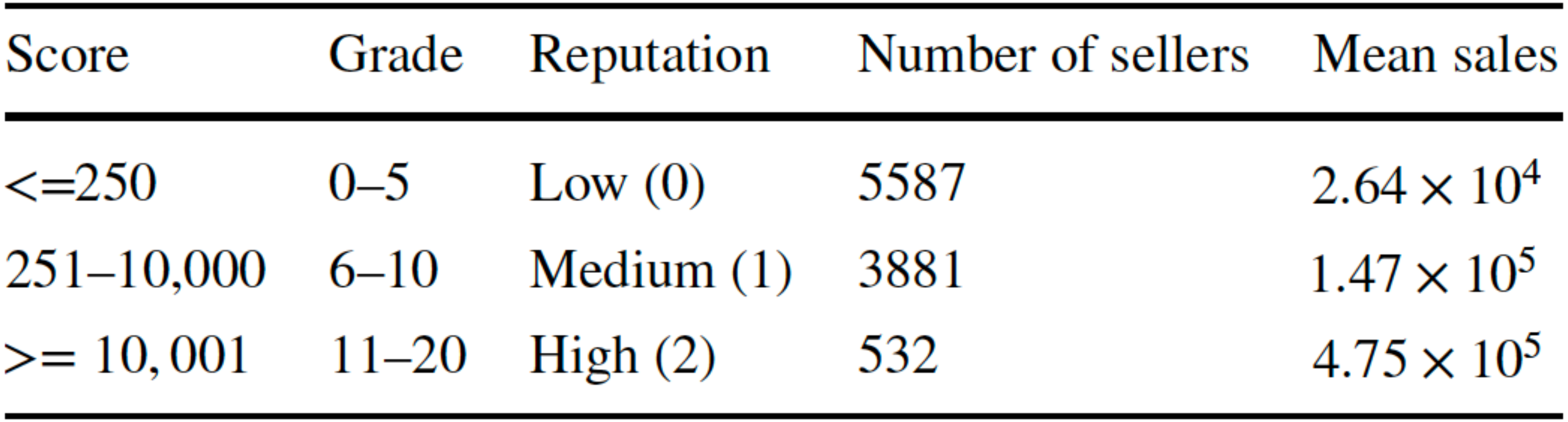
Real example:

- Apply the proposed methodology to the Taobao dataset collected from Taobao.com, China’s largest e-commerce platform.
- Taobao offers a feedback system to reduce the likelihood of fraud and encourage trust based on reputation.
- The website adopts a similar reputation rating system in which the cumulative rating score is then categorized into twenty grades from 0 to 20.
- Based on this, the authors define the reputation variable as follows

0-5 Low, 6-10 Medium, and 11-20 High

- The dataset we get is a simple random sample from the original database with a sample size of ten thousand.

Table 2 Seller’s reputation and average sales



Score	Grade	Reputation	Number of sellers	Mean sales
<=250	0–5	Low (0)	5587	2.64×10^4
251–10,000	6–10	Medium (1)	3881	1.47×10^5
>= 10, 001	11–20	High (2)	532	4.75×10^5

- For each seller from June 2011 to December 2011 data were collected.
- In this analysis, the authors are interested in examining the causal effect of a seller’s rating (reputation) on sales (gross revenue).
- There are 13 potential confounders related to the seller’s characteristics like the seller’s age, the seller’s gender, etc.

Summary of the results

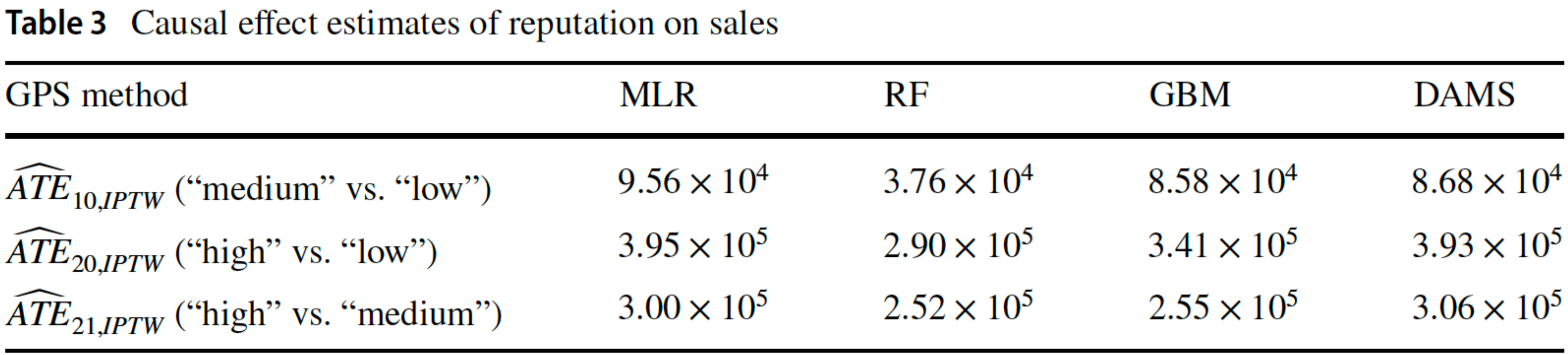


Table 3 Causal effect estimates of reputation on sales

GPS method	MLR	RF	GBM	DAMS
$\widehat{ATE}_{10,IPTW}$ (“medium” vs. “low”)	9.56×10^4	3.76×10^4	8.58×10^4	8.68×10^4
$\widehat{ATE}_{20,IPTW}$ (“high” vs. “low”)	3.95×10^5	2.90×10^5	3.41×10^5	3.93×10^5
$\widehat{ATE}_{21,IPTW}$ (“high” vs. “medium”)	3.00×10^5	2.52×10^5	2.55×10^5	3.06×10^5

- RF seems to work quite differently and the rest of the three methods yield similar results.
- We can conclude that sellers with “medium” reputation has an increase of around 9×10^4 Yuan in half-year sales, compared to “low” reputation.

Discussion

- From the simulation studies, we can say that using MLR to estimate the generalized propensity scores can result in extreme weights that in turn result in more bias.
- On the other hand, GBM, DAMS, and RF tend to be more stable across different levels of complexity in the relationship between the treatment assignment and the covariates.
- DAMS performed the best out of the four propensity score estimation methods in combination with IPTW.
- In conclusion, we can recommend machine learning methods for propensity score estimation in multi-level treatment settings.
- In real applications, the true propensity score model is never known for observational studies. Researchers may apply different algorithms and see how different the results are.

Thank you