Eli Lauwers - Causality Project

2022-07-24

Q1: Logistic Regression outcome model

Following estimates were obtained.

	Risk Difference	Relative Risk
Manual	0.132	1.206
stdReg	0.132	1.206

Argumentation: Using backwards step logistic regression with cutoff of .1, I built an outcome model retaining following predictors: `qsmk` (the treatment), `dbp`, `sex`, `age`, `wt71`, `smokeintensity`, `active`. I used manual G-computation to predict outcomes $P(Y^1 = 1)$ and $P(Y^0 = 1)$. More specifically, I used the fitted model to predict the outcomes for every person in the dataset while manually fixing the `qsmk` value to 0 or 1. Next, I used the `stdReg` package to cross check results.

Q2: Advantages of causal machine learning

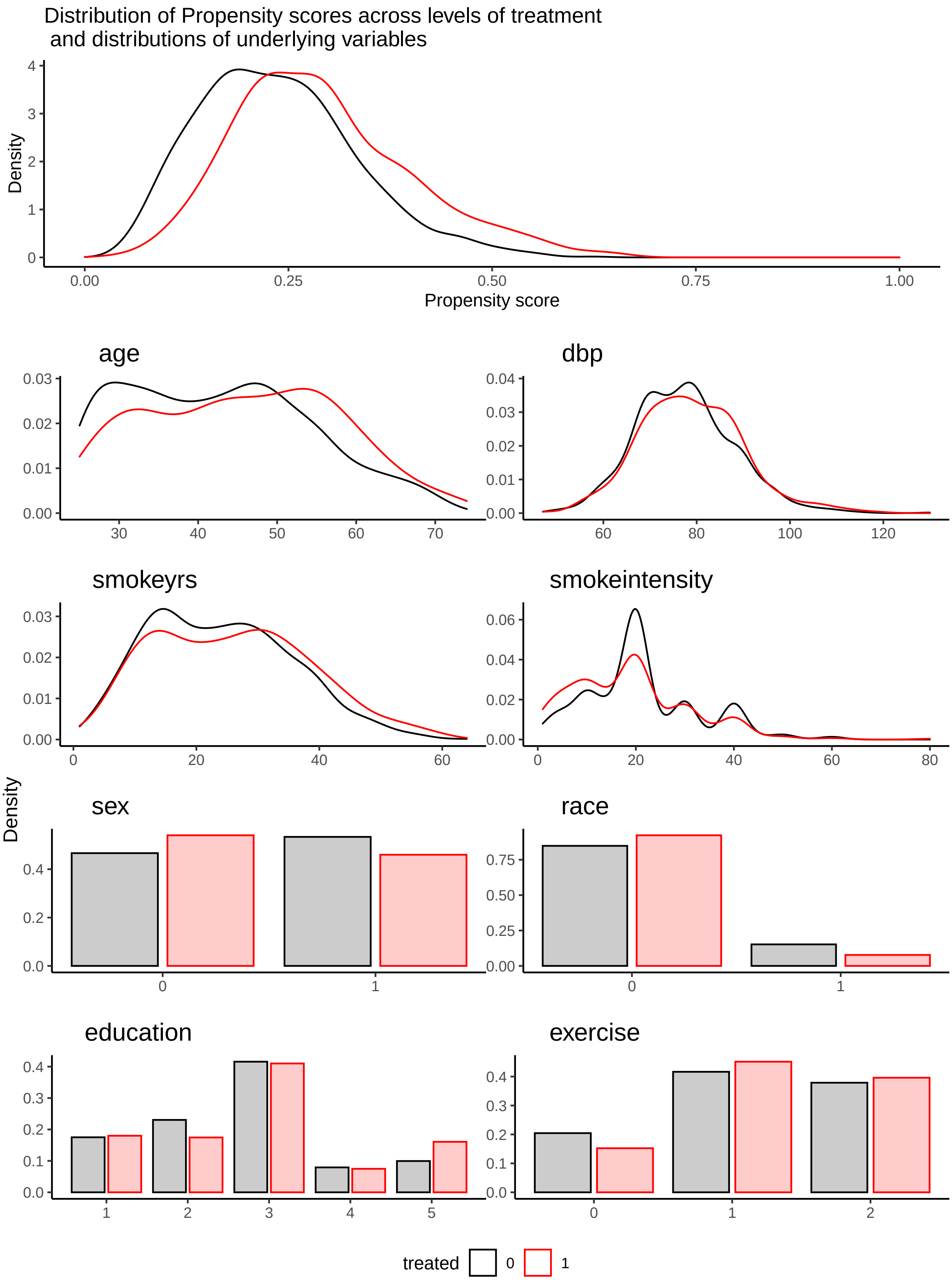
Answer: two advantage of causal machine learning versus more classic statistical techniques are:

1. It allows for the use of data-adaptive algorithms (e.g.: machine learning in the outcome and treatment model) without sacrificing interpretability of estimates.
2. The estimators are *double robust*, which means that estimates are *correct* (efficient and consistent) when either one of the outcome model or propensity score model is correctly specified.

Q3: Estimate propensity scores

Answer: There is significant overlap of propensity scores across treatment groups. However, some intergroup differences can be found in the propensity scores as well as in the underlying variables. As a result, some bias will be introduced due to extropolation in areas of the propensity scores where one treatment group largely dominates the other. Propensity score distributions for both treatment groups seem fairly identical, although the treated have on average higher propensity scores than the non-treated. It follows that treated people are more likely to have been treated, and both groups are not entirely (conditionally) exchangeable. However, **since both distributions are so similar, I don't think the issues of non-overlap will have high impact on estimates.**

Argumentation: I fitted a logistic regression model using backwards step regression with the treatment (`qsmk`) as the dependent variable. I used a cutoff of .1 and retained following predictors: `dbp`, `sex`, `age`, `race`, `education`, `smokeintensity`, `smokeyrs`, `exercise`. Next, I extracted the fitted values from the model object as the propensity scores


Distribution of Propensity scores across levels of treatment
and distributions of underlying variables
Density
0
1
2
3
4
0.00
0.25
0.50
0.75
1.00
Propensity score
age
0.00
0.01
0.02
0.03
30
40
50
60
70
dbp
0.00
0.01
0.02
0.03
0.04
60
80
100
120
smokeyrs
0.00
0.01
0.02
0.03
0
20
40
60
smokeintensity
0.00
0.02
0.04
0.06
0
20
40
60
80
Density
sex
0.0
0.2
0.4
0
1
race
0.00
0.25
0.50
0.75
0
1
education
0.0
0.1
0.2
0.3
0.4
1
2
3
4
5
exercise
0.0
0.1
0.2
0.3
0.4
0
1
2
treated
0
1

Q4: $\widehat{ATE}_{TMLE}$ for 3 modelling cases

Answer: following estimates were obtained:

Technique	$\widehat{ATE}_{TMLE}$	SE	95% CI
Tutorial	0.130	0.027	[0.076, 0.184]
Superlearner Default	0.138	0.025	[0.088, 0.188]
Superlearner Custom	0.140	0.023	[0.094, 0.185]

Note: for questions 4 up to 7, I used the same set of covariates in the model building exercise for the treatment and outcome model. More specific, I selected (via double selection) the subset of covariates which were either associated with outcome or with treatment using a cutoff of .10. Following variables were retained by means of double selection:`qsmk` (the treatment), `dbp`, `sex`, `age`, `wt71`, `race`, `education`, `smokeintensity`, `smokeyrs`, and `exercise`

Interpretation: Although there are no significant differences between estimates - every estimate lies in the confidence intervals of the others - , two trends seem to appear. First, the $\widehat{ATE}_{TMLE}$ is smallest for the tutorial modelling case, which is based on simple logistic regression models for treatment and outcome. The $\widehat{ATE}_{TMLE}$ grows larger with the default SuperLearner technique and grows even further by use of the custom SuperLearner libraries.

The second finding is the fact that the standard error gets smaller from the tutorial estimate to the SuperLearner estimate with custom libraries.

Both findings combined, it seems as though the prediction of either outcome or treatment is improved by exchanging (simple) logistic linear models for more intricate and data-adaptive measures. As the constructed $\widehat{ATE}_{TMLE}$ is based on a double robust estimator, the estimate is consistent and efficient when either the treatment model or outcome model is specified correctly.

Q5: $P_{AIPTW}(Y^a = 1)$ for 3 modelling cases

Answer: Following estimates were obtained

Technique	$P(Y^1 = 1)$	SE	95% CI	$P(Y^0 = 1)$	SE	95% CI
Tutorial	0.768	0.024	[0.721, 0.815]	0.638	0.015	[0.609, 0.666]
Superlearner Default	0.771	0.011	[0.750, 0.793]	0.638	0.039	[0.561, 0.715]
Superlearner Custom	0.771	0.010	[0.750, 0.791]	0.643	0.038	[0.567, 0.718]

Answer: Looking at the results table, multiple trends emerge. Firstly, for every used technique, estimates for $P(Y^1 = 1)$ are significantly larger than than estimates for $P(Y^0 = 1)$. If every person in the population quit smoking, then more people would experience weight gain than in a situation where every person continued smoking.

A second finding is the fact that standard error for $P(Y^1 = 1)$ goes down by switching simple linear logistic models for the superlearner approach, while the opposite is true for $P(Y^0 = 1)$.

A third finding is that standard errors within an estimate are fairly similar between the default and custom implementation of SuperLearner but are dissimilar between the usage of SuperLearner and the more simple logistic regression approach.

Q6: $\widehat{ATE}_{TMLE}$ versus $\widehat{ATE}_{AIPTW}$

Answer:

Based on AIPTW

Technique	$\widehat{ATE}_{AIPTW}$	SE	95% CI
Tutorial	0.130	0.027	[0.076, 0.184]
Superlearner Default	0.133	0.040	[0.054, 0.213]
Superlearner Custom	0.128	0.039	[0.051, 0.205]

Based on TMLE (question 4)

Technique	$\widehat{ATE}_{TMLE}$	SE	95% CI
Tutorial	0.130	0.027	[0.076, 0.184]
Superlearner Default	0.138	0.025	[0.088, 0.188]
Superlearner Custom	0.140	0.023	[0.094, 0.185]

Interpretation: When looking at the results table, estimates for $\widehat{ATE}$ seem fairly similar across estimators and estimation techniques. Point estimates range from 13% and 14%, with no remarkable differences between estimators or estimator techniques.

A first notable remark is that estimates and standard errors are identical for the tutorial versions of both estimators. In the tutorial, both treatment and outcome models are based on simple logistic regression.

A second finding is the difference in standard errors. For $\widehat{ATE}_{AIPTW}$, the standard errors range from .027 to .040. For the other estimate ($\widehat{ATE}_{AIPTW}$), standard errors lay somewhat lower, ranging from .023 to .027.

A third notion is found when comparing estimates for the tutorial technique against the SuperLearner approach between estimators. For $\widehat{ATE}_{AIPTW}$, Estimates for the SuperLearner techniques are less precise than the estimate for the tutorial-approach. For $\widehat{ATE}_{TMLE}$, this notion is reversed, as standard errors for the SuperLearner estimates are smaller than those for the tutorial estimate. However, inter-estimator technique precision differences for $\widehat{ATE}_{TMLE}$ are smaller than those for $\widehat{ATE}_{AIPTW}$.

Q7: $P_{TMLE}(Y^a = 1)$ versus $P_{AIPTW}(Y^a = 1)$

Answer:

Based on TMLE

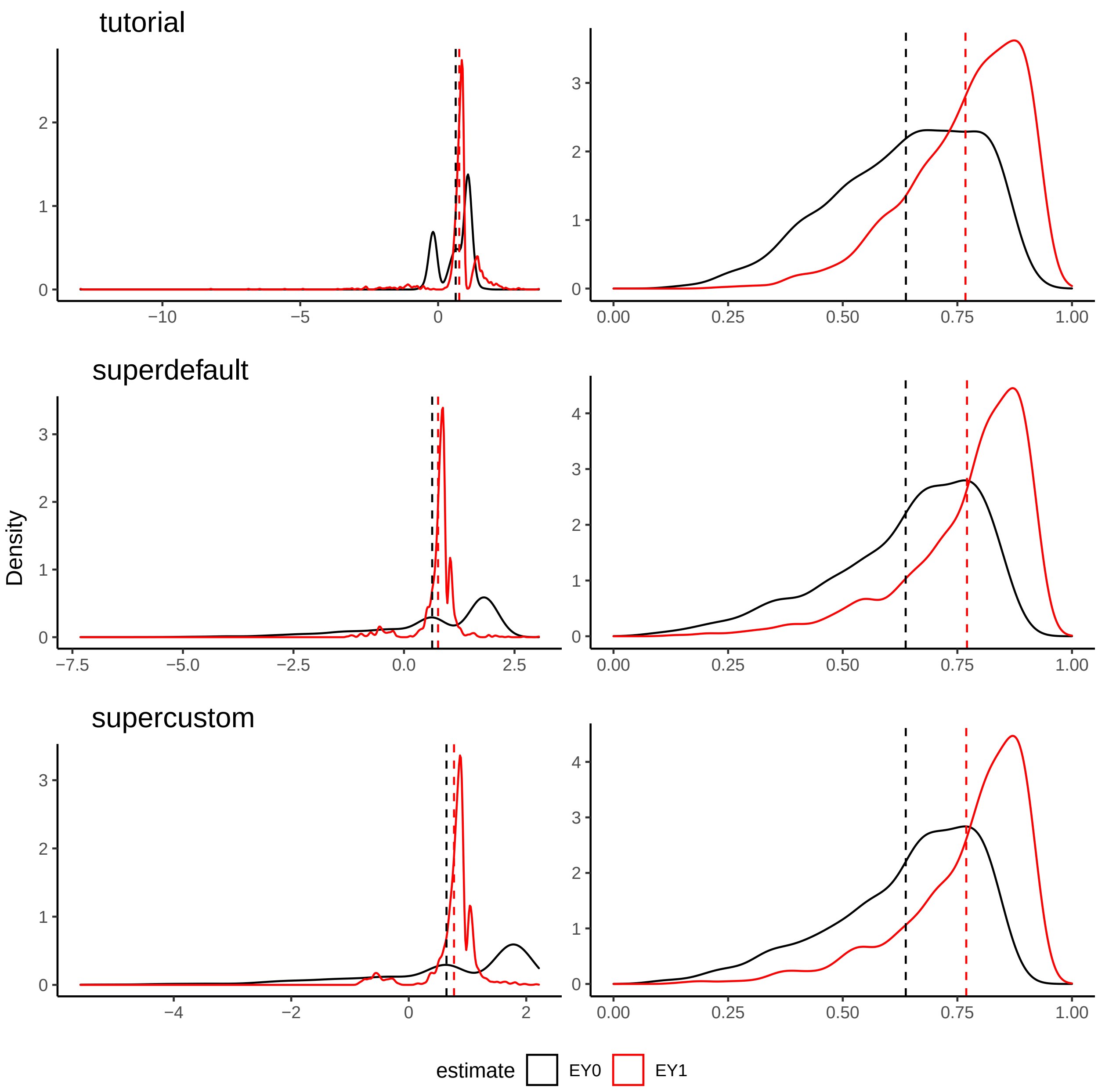
Technique	$P(Y^1 = 1)$	SE	95% CI	$P(Y^0 = 1)$	SE	95% CI
Superlearner Custom	0.768	0.003	[0.761, 0.774]	0.638	0.004	[0.630, 0.646]
Superlearner Custom	0.771	0.004	[0.764, 0.778]	0.637	0.004	[0.629, 0.646]
Superlearner Custom	0.769	0.004	[0.762, 0.776]	0.637	0.004	[0.629, 0.646]

Based on AIPTW (Question 5)

Technique	$P(Y^1 = 1)$	SE	95% CI	$P(Y^0 = 1)$	SE	95% CI
Tutorial	0.768	0.024	[0.721, 0.815]	0.638	0.015	[0.609, 0.666]
Superlearner Default	0.771	0.011	[0.750, 0.793]	0.638	0.039	[0.561, 0.715]
Superlearner Custom	0.771	0.010	[0.750, 0.791]	0.643	0.038	[0.567, 0.718]

Interpretation: In question 6, the estimate for $\widehat{ATE}_{TMLE}$ was on average more precise than the estimate for $\widehat{ATE}_{AIPTW}$. Comparing estimates for $P(Y^a = 1)$ for both estimating techniques, point estimates do not seem to vary greatly across estimates. However, inter-estimator technique precision of estimates $P(Y^a = 1)$ differ greatly. Whereas standard errors for $P_{AIPTW}(Y^a = 1)$ range from .010 to .040, the standard errors for both estimates $P_{TMLE}(Y^a = 1)$ are equal to .004, making TMLE more precise as an estimator technique than AIPTW in this particular use case. A remark to this finding is that the $P_{TMLE}(Y^a = 1)$ were - as the project requirements require - only based on the SuperLearner technique using the custom library set. It is unsure if the standard error for $P_{TMLE}(Y^a = 1)$ would be more precise across estimates.

Distributions for $P(Y^a = 1)$


tutorial
superdefault
supercustom
Density
estimate
EY0
EY1