Eli Lauwers - Causality Project

2022-07-24

Background and Preparation

Read the dataset and modify it as follows:

In our analysis of this observational study we will aim to evaluate the marginal effect of quitting smoking (the information is held in the variable `qsmk`) on the weight gain between years 1971 and 1982 (the information is held in the variable `wt82_71_bin`, which is 1, in case of weight gain and 0 otherwise). The tutorial [1] may be a helpful reference for your analysis. Throughout, you may assume that the dataset contains a collection of covariates that is sufficient to control for confounding of the effect of quitting smoking on the risk of weight gain.

Q1: Logistic Regression outcome model

Question: Estimate the marginal effect of quitting smoking on the risk of weight gain as a risk difference and a relative risk using a logistic regression outcome model. There is no need to be very exhaustive in building a model. You may for instance consider using a standard model choice, and make use of automated procedures (e.g. step in R) to select variables, if needed. More thorough model building will be considered later using causal machine learning methods.

answer: 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 predictors `qsmk` (the treatment), `dbp`, `sex`, `age`, `wt71`, `smokeintensity`, `active`. I used manual G-computation to predict counterfactual outcomes. 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. I also used the `stdReg` package to cross check results.

Q2: Advantages of causal machine learning

Question: Briefly discuss what you view as possible advantages of causal machine learning approaches (like those considered from question 4 onwards), relative to more standard statistical analyses as in question 1. List 2 possible advantages.

Answer:

1. It allows the use of data-adaptive algorithms like machine learning without sacrificing interpretability.
2. The estimators are *double robust*, which means that estimates are correct when either one of the outcome model or propensity score model is correctly specified.

Q3: Propensity scores

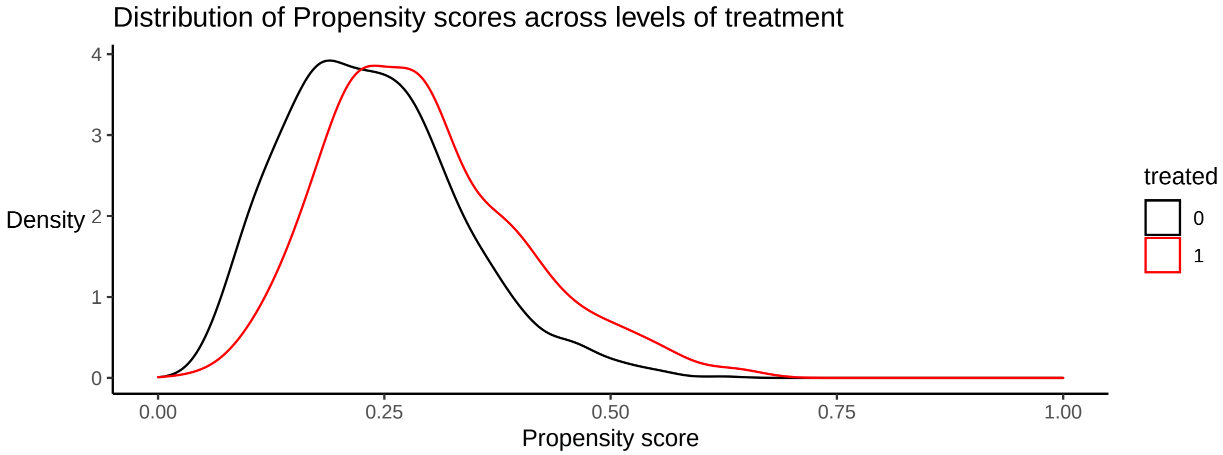
Question: Use logistic regression to compute the propensity score for quitting smoking. There is no need to be very exhaustive in building a model. You may for instance consider using a standard model choice, and make use of automated procedures (e.g. step in R) to select variables, if needed. More thorough model building will be considered later using causal machine learning methods.

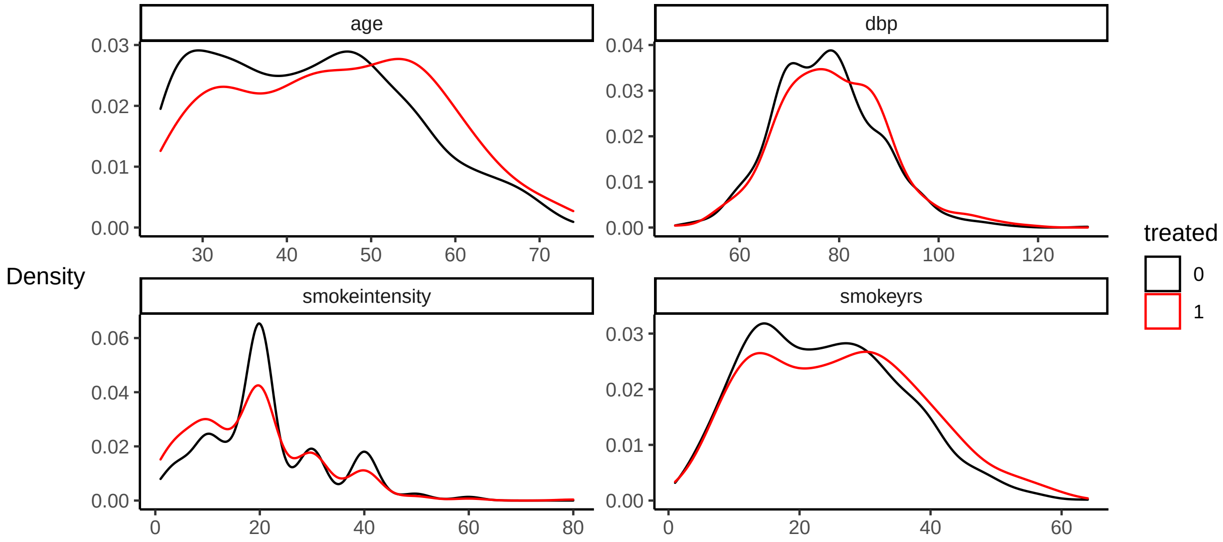
Discuss the balance of covariates, distribution of propensity scores in both exposure groups and the overlap. Discuss if, based on the obtained results, you foresee difficulties in inferring the effect of quitting smoking on the risk of weight gain.

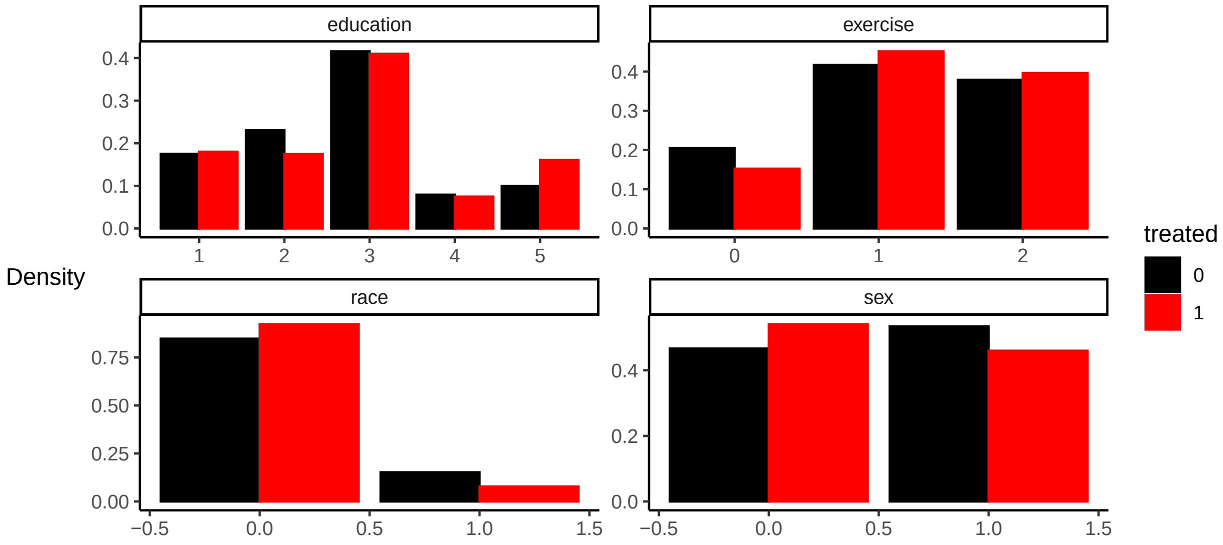
Answer: Although there is significant overlap between both treatment groups in terms of propensity scores, There still are some inter-group differences. More specific, the distribution of the propensity scores for the treated group seems similar to the distribution for the untreated, although shifted to the right. It follows that treated people are more likely to have been treated, and both groups are not entirely (conditionally) exchangeable. It also follows that the final models will use extrapolation in areas of the propensity score where either group is not really present.

However, as the overlap between treatment groups is as high, I don't think the non-overlapping areas will have much impact on the final 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






education
exercise
race
sex
Density
treated
0
1
0.0
0.1
0.2
0.3
0.4
1
2
3
4
5
0
1
2
0.00
0.25
0.50
0.75
−0.5
0.0
0.5
1.0
1.5

Q4: TMLE for 3 modelling cases

Question: Calculate the targeted maximum likelihood estimate (TMLE) of the marginal effect of quitting smoking on the risk of weight gain (on the risk difference scale). Consider 3 cases in terms of modelling the propensity score and the risk of weight gain:

- logistic regression models obtained above;
- SuperLearner with default options for the outcome regression and propensity score models;
- SuperLearner with the following selection of the models for the outcome regression and propensity score models.

```
SL.library <- c("SL.glm","SL.step","SL.step.interaction", "SL.glm.interaction", "SL.gam",
"SL.randomForest", "SL.rpart")
```

For this, you can either use the `tmle` function in R, or follow the steps of the tutorial. Estimate the standard errors and 95% confidence intervals for the obtained estimates (there is no need for using the bootstrap). Compare the results for different modelling choices

Answer:

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]

Argumentation

Q5: AIPTW for 3 modelling cases

Question: Use augmented inverse probability of treatment weighting to estimate the (counterfactual) risk of weight gain, $P(Y^1 = 1)$, that would be seen if all were to quit smoking, as well as when no one were to quit smoking ($P(Y^0 = 1)$). As in the previous exercise, consider 3 cases in terms of modelling the propensity score and the risk of weight gain:

- logistic regression models obtained above;
- SuperLearner with default options for the outcome regression and propensity score models;
- SuperLearner with the following selection of the models for the outcome regression and propensity score models.

```
SL.library <- c("SL.glm","SL.step","SL.step.interaction","SL.glm.interaction","SL.gam","SL.randomForest
```

Estimate the standard errors and 95 confidence intervals for the obtained estimates (there is no need for using the bootstrap). Compare the results for different modelling choices.

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:

Q6: ATE for results AIPTW and TMLE

Question: An AIPTW estimate for the marginal effect of quitting smoking on the risk of weight gain (on the risk difference scale) is readily obtained as the difference between the estimates for $P(Y^1 = 1)$ and $P(Y^0 = 1)$, obtained in question 5. Report this difference along with a standard error (there is no need for using the bootstrap), and compare the result with the TMLE results from question 4.

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]

Q7: Estimates for $P(Y^a = 1)$ for AIPTW and TMLE

Question: Estimate $P(Y^1 = 1)$ and $P(Y^0 = 1)$ using TMLE with SuperLearner (with user-selected libraries as above); see the tutorial for how to do this (as the tmle function does not readily do this). Estimate the standard errors and 95% confidence intervals for the targeted estimates. Compare the results with those obtained via AIPTW in question 5

Answer:

Based on TMLE

Technique	$P(Y^1 = 1)$	SE	95% CI	$P(Y^0 = 1)$	SE	95% CI
Superlearner Custom	0.768	0.004	[0.761, 0.775]	0.638	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]