Eli Lauwers - Causality Project

2022-07-24

Background

Copy background

Preparation

Read the dataset and modify it as follows:

(code available in the .Rmd file)

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.

1 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: I used manual g-computation and a logistic outcome model (with predictors `qsmk`, `dbp`, `sex`, `age`, `wt71`, `smokeintensity`, `active`) to predict counterfactual outcomes. following results were obtained

- Risk Difference: **ADD** (p1 - p0)
- Relative risk: **ADD** (p1 / p0)

I also used the `stdReg` package to cross check results.

Argumentation:

For the manual approach, I used the backwards step model to predict counterfactual outcomes for every instance in the dataset, while manually setting the treatment (`qsmk`) variable to a certain level. I only used predictions for instances without any missing data.Next, I averaged the results and calculated the effect measures.

For information on the custom `step_backwards`-function, please see the appendix. In short, it uses a `while`-loop which refits the model until (1) every coefficient has a p value under the cut-off (.5) or (2) the treatment variable (`qsmk`) has the highest p value.

```r
source("step_backwards.R")
# verbose 0 => do not print anything
predictors.all = names(nhefs.nmv)[names(nhefs.nmv) != "wt82_71_bin"]
model = step_backwards(data = nhefs.nmv, outcome = "wt82_71_bin", treatment = "qsmk",
    predictors = predictors.all, cutoff = 0.1, verbose = 1)
```

```
Predictors in backwards procedure for wt82_71_bin:
   Kept: qsmk, dbp, sex, age, wt71
   Removed: race, smokeyrs, education, tax71_82, sbp, exercise, smokeintensity, active
```

```r
# Using G computation
p0 = mean(predict(model, na.omit(mutate(nhefs.nmv, qsmk = 0)), type = "response"))
p1 = mean(predict(model, na.omit(mutate(nhefs.nmv, qsmk = 1)), type = "response"))
rd.g = p1 - p0
rr.g = p1/p0

# crosscheck results using stdReg
model.std = stdReg::stdGlm(model, data = nhefs.nmv, X = "qsmk", x = c(0, 1))
# difference estimate
model.std.diff = summary(model.std, contrast = "difference", reference = 0)
rd.stdreg = model.std.diff$est.table[2, "Estimate"]
# Ratio estimate
model.std.ratio = summary(model.std, contrast = "ratio", reference = 0)
```

	Risk Difference	Relative Risk
Manual	0.132	1.206
stdReg	0.132	1.206

2 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:

3 Propensity scores

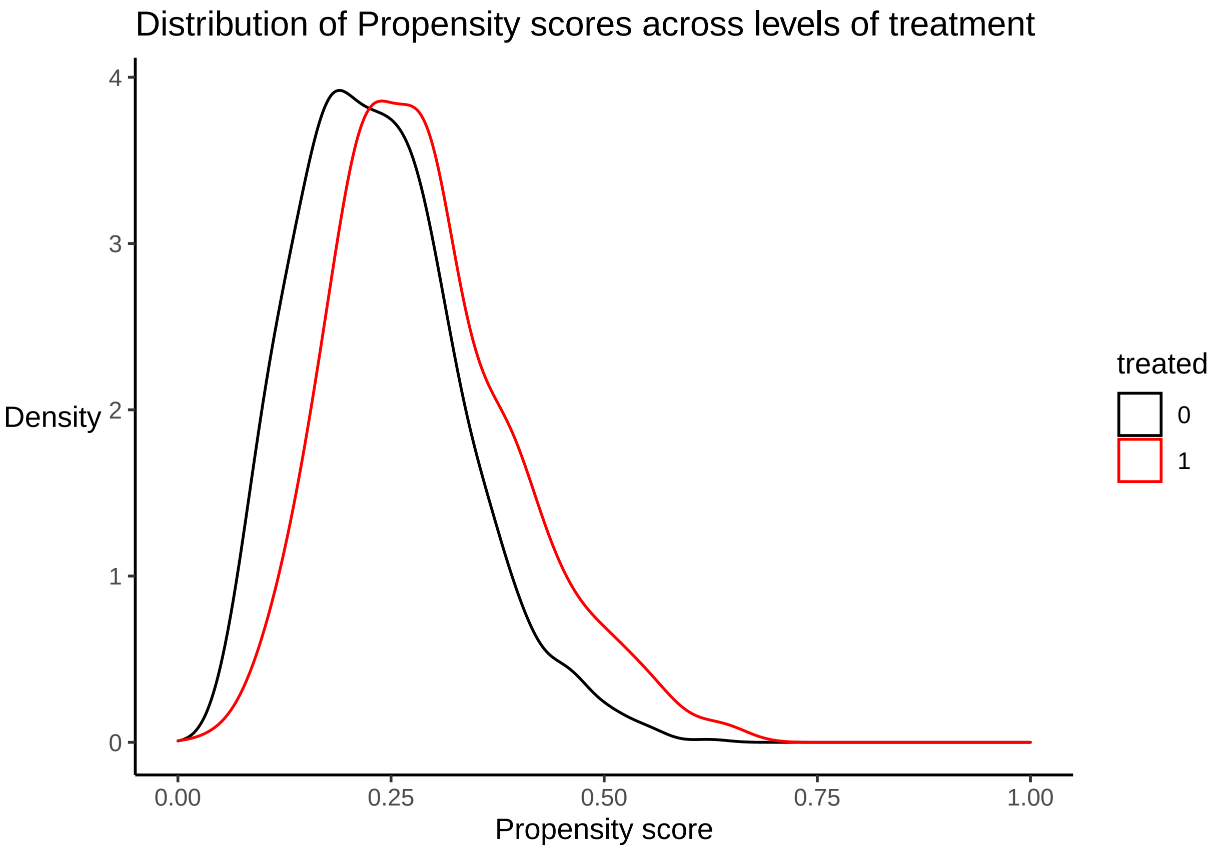
Question: Use logistic regression to compute the propensity score for quitting smoking. There is no need to be very exhaustive in building a model. You may for instance consider using a standard model choice, and make use of automated procedures (e.g. step in R) to select variables, if needed. More thorough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

Argumentation:

```r
ps.predictors = names(nhefs.nmv)[!names(nhefs.nmv) %in% c("wt82_71_bin", "qsmk")]
model = step_backwards(data = nhefs.nmv, outcome = "qsmk", predictors = ps.predictors,
    cutoff = 0.1, verbose = 0)
ps.scores = fitted(model)
```



4 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:

Argumentation

```
# Step 1: the outcome model
preds = names(nhefs.nmv)[names(nhefs.nmv) != "wt82_71_bin"]
m = step_backwards(nhefs.nmv, outcome = "wt82_71_bin", treatment = "qsmk", predictors = preds,
    cutoff = 0.1, verbose = 1)
```

```
Predictors in backwards procedure for wt82_71_bin:
   Kept: qsmk, dbp, sex, age, wt71
   Removed: race, smokeyrs, education, tax71_82, sbp, exercise, smokeintensity, active
```

```
QAW = predict(m, type = "response")
Q1W = predict(m, newdata = mutate(nhefs.nmv, qsmk = 1), type = "response")
Q0W = predict(m, newdata = mutate(nhefs.nmv, qsmk = 0), type = "response")
mean(Q1W - Q0W)
```

```
[1] 0.1318421
```

```
# Step 2: propensity score
preds = names(nhefs.nmv)[!names(nhefs.nmv) %in% c("qsmk")]
psm = step_backwards(nhefs.nmv, outcome = "qsmk", predictors = preds, cutoff = 0.1,
    verbose = 1)
```

```
Predictors in backwards procedure for qsmk:
   Kept: dbp, sex, age, race, education, wt71, smokeintensity, smokeyrs, exercise, wt82_71_bin
   Removed: sbp, tax71_82, active
```

```
gW = predict(psm, type = "response")
# step 3: The clever covariate
H1W = nhefs.nmv$qsmk/gW
H0W = (1 - nhefs.nmv$qsmk)/(1 - gW)
# fluctuation parameter
```

```
epsilon = coef(glm(nhefs.nmv$wt82_71_bin ~ -1 + H0W + H1W + offset(qlogis(QAW)),
    family = binomial))
# Step 4: update of original outcome model
Q0W_1 = plogis(qlogis(Q0W) + epsilon[1]/(1 - gW))
Q1W_1 = plogis(qlogis(Q1W) + epsilon[2]/gW)
ATEtmle1 = mean(Q1W_1 - Q0W_1)
EY1tmle1 = mean(Q1W_1)
EY0tmle1 = mean(Q0W_1)
MORtmle1 = (EY1tmle1 * (1 - EY0tmle1))/((1 - EY1tmle1) * EY0tmle1)
# Step 6: Inference and CI ATE efficient influence curve (EIC)
D1 <- nhefs.nmv$qsmk/gW * (nhefs.nmv$wt82_71_bin - Q1W_1) + Q1W_1 - EY1tmle1
D0 <- (1 - nhefs.nmv$qsmk)/(1 - gW) * (nhefs.nmv$wt82_71_bin - Q0W_1) + Q0W_1 - EY0tmle1
EIC <- D1 - D0
# ATE variance
n <- nrow(nhefs.nmv)
varHat.IC <- var(EIC)/n
# ATE 95%CI
ATEtmle1_CI <- c(ATEtmle1 - 1.96 * sqrt(varHat.IC), ATEtmle1 + 1.96 * sqrt(varHat.IC))
ATEtmle1
```

```
[1] 0.001191491
```

```
ATEtmle1_CI
```

```
[1] -0.05444343  0.05682641
```

```
# ATEtmle1_CI(95%CI): 22.1% (15.1, 29.0) MOR EIC
EIC <- (1 - EY0tmle1)/EY0tmle1/(1 - EY1tmle1)^2 * D1 - EY1tmle1/(1 - EY1tmle1)/EY0tmle1^2 *
    D0
varHat.IC <- var(EIC)/n
# MOR 95%CI
MORtmle1_CI <- c(MORtmle1 - 1.96 * sqrt(varHat.IC), MORtmle1 + 1.96 * sqrt(varHat.IC))
MORtmle1
```

```
[1] 1.00544
```

```
MORtmle1_CI
```

```
[1] 0.7506145 1.2602648
```