Project: Causal Machine Learning

Causal inference and missing data

We use data for a population of people who smoked in 1971 who participated in the National Health and Nutrition Examination Survey Data I Epidemiologic Follow-up Study (NHEFS). The data file and the codebook with variable descriptions can be found here: `https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/`. Furthermore, a detailed description of the NHEFS can be found at `https://wwwn.cdc.gov/nchs/nhanes/nhefs/`.

Read the dataset and modify it as follows:

```
library("readxl")
nhefs <- read_excel("../nhefs.xls")

# Ignore patients with missing outcome
nhefs.nmv <- nhefs[which(!is.na(nhefs$wt82_71)),]

# Dichotomize the outcome variable
nhefs.nmv$wt82_71_bin <- ifelse(nhefs.nmv$wt82_71 > 0, 1, 0)

# Select reduced dataset
nhefs.nmv <- nhefs.nmv[,(names(nhefs.nmv) %in% c("wt82_71_bin", "qsmk", "sex",
"dbp", "race", "sbp", "age", "education", "smokeintensity", "smokeyrs", "tax71_82",
"exercise", "active", "wt71"))]
```

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. 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.

2. 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.

3. 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.

4. 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

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

     for the outcome regression and propensity score models.

   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.

5. 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

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

     for the outcome regression and propensity score models.

   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.

6. 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.

7. 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.

Make a **concise, but complete** report describing what analyses you performed and what are your conclusions to the various questions. There is no need to be extensive in your report. For instance, there is no need to repeat the theory that was used, as you may assume that the reader is familiar with the procedures that you used. However, you should carefully interpret the reported estimates and make sure that their meaning is clear and not obscure. Where possible, try to explain the differences observed between estimates obtained via different methods and give insight to the reader. **Make clear what estimates you would consider as final if you could only choose on result to report.** Overall, I expect that 2 or 3 pages should suffice; 5 pages is a maximum.

You can work on this project in groups of at most 3 people. It is recommended that you split the work, **and then give feedback on each other's report**. If you split the work, the project should not take too much work, but otherwise it may. A recommended split is one person focus on question 1 and 4, one on 2 and 5, one on 3 and 6, and all on 7. If one of you is comfortable in making question 7, a more efficient split is one person focus on question 1 and 4, one on 2, 5 and 6 (6 is a short question) and one on 3 and 7. One report per group must be returned no later than no later than May 1, together with structured software code (i.e. a script file). All the files that you submit, must carry the family name of one of the members in the group.

The amount of work is highest for questions 4, 5 and 7, little for question 1, 2 and 6, and medium for question 3.

Make sure your report briefly mentions who did what. Two thirds of the score will be on your own contribution and one third on the project as a whole (as indeed you are also responsible for the entirety of the project).

References

[1] Luque-Fernandez MA, Schomaker M, Rachet B, Schnitzer ME. Targeted maximum likelihood estimation for a binary treatment: A tutorial. Statistics in Medicine. 2018;37:2530-2546.