Problem Set 1 - *Potential Outcome Model*

Due 17. April 2020

Data set: The *National Health Interview Survey (NHIS)*[1] data has been collected on U.S. households since 1957; it covers a broad range of health related topics from medical conditions, health insurance and number of doctors visits to measures of physical activity. Here we focus on indicators relevant for the POM framework; in particular, we will compare the health status of hospitalized and nonhospitalized individuals in 2018. For this purpose, we use answers to the survey question "During the past 12 months, has the respondent been hospitalized overnight?" with potential answers "yes" and "no" which we code as 1 and 0. Further we consider answers to the questions "Would you say your health in general is excellent, very good, good, fair, poor?" where answers are coded as 1 for poor health up to 5 for excellent health. The survey also collects data on relevant characteristics as sex, age, level of education, hours worked last week and total earnings.

Problem 1.1.

i.) Open a `Jupyter Notebook` and import the data set `nhis-initial.xslx` —we recommend using the software presented in class (e.g. `pandas`, `matplotlib`, etc.). Try to think of ways to answer the following questions: *Are there more females or males? Are there more inidividuals who hold a degree or not?*. Now try to relate individual characteristics to the hospitalization status. *Are high or low earners healthier? Are more old or young people hospitalized?*

ii.) Compute the average health status of hospitalized and nonhospitalized individuals. *Who is healthier on average? What could be a reason for this bias?*

iii.) Adjust the data set for the POM framework (as seen in the lecture), with health status as the outcome and hospitalization as the treatment status (Hint: rename and drop columns of the data frame).

iv.) Compute the naive estimate for the *average treatment effect* (ATE).

Problem 1.2.

i.) As we've seen in the lecture, in reality we can only ever observe one counterfactual; however, when simulating data we can bypass this problem. The (simulated) data set `nhis-simulated.xslx` contains counterfactual outcomes, i.e. outcomes under the control for individuals who were assigned to the treatment group and vice versa. Derive and compute the average outcomes in the two observable and two nonobservable states; Design them in a similar way as table 2.3 in (Morgan & Winship, 2014).

From here on assume that 5% of the population take the treatment.

ii.) Derive and explain formula 2.12 from Morgan and Winship (2014) for the naive estimator as a decomposition of true ATE, baseline bias, and differential treatment effect bias (do this with pen and paper).

iii.) Compute the naive estimate and true value of the ATE for the simulated data. *Is the naive estimator upwardly or downwardly biased?* Calculate the baseline bias and differential treatment effect bias. *How could we interpret these biases in our framework of health status of hospitalized and nonhospitalized respondents?*

iv.) Which assumptions must hold (on the data) such that the naive estimator is an *unbiased* and *consistent* estimator for the ATE?

[1] https://www.cdc.gov/nchs/nhis/index.htm

References

Angrist, J. D., & Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion.* Princeton, NJ, USA: Princeton University Press.

Morgan, S. L., & Winship, C. (2014). *Counterfactuals and causal inference.* West Nyack: Cambridge University Press.