Potential Outcome Model*

– Problem Set 1 –

The `National Health Interview Survey (NHIS)` data is collected on U.S. households since 1957; it covers a broad range of health-related topics from medical conditions, health insurance, and the number of doctor 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 non-hospitalized 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 responses 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.

Task A

(1) Open a `Jupyter Notebook` and import the data set `nhis-initial.xslx` (available at `https://bit.ly/nhis-initial`) — 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 individuals who hold a degree or not?*. Now try to relate individual characteristics to the hospitalization status. *Are high or low earners/old or young people more often hospitalized?*

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

*We are indebted to Liudmila Kiseleva and Tim Mensinger for the design of the problem set.

(3) 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).

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

Task B

(1) 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` (available at `https://bit.ly/nhis-simulated`) contains counterfactual outcomes, i.e., outcomes under control for individuals assigned to the treatment group and vice versa. Derive and compute the average outcomes in the two observable and two unobservables states; Design them in a similar way as Table 2.3 in Morgan & Winship (2014).

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

(2) Derive and explain formula 2.12 from Morgan & 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).

(3) 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 non-hospitalized respondents?*

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

References

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

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

National Health Interview Survey. (2018). National Center for Health Statistics. Retrieved from: https://www.cdc.gov/nchs/nhis/index.htm.