Assignment 2
Causal Inference

Deadline January 24th, 2025

Handout for the *Bayesian Networks and Causal Inference* lecture, November 25th, 2024

Johannes Textor, Moabi Mokhoro

Objectives of This Exercise

1. Perform a causal inference analysis on a real dataset.
2. Interpret and report the results of your causal inference analysis using appropriate language and graphics.

Tasks

There are two options for this assignment:

- Take the DAG from assignment 1. Pick a focal relationship (i.e., a defined exposure and a defined outcome variable). Perform *n* causal inference analyses from the list below on this focal relationship, where *n* is the size of your team.
- Take the DAG from assignment 1. Pick a focal relationship (i.e., a defined exposure and a defined outcome variable). Use a structure learning algorithm to automatically learn a structure from the dataset; this will need to be a structure learning algorithm where you can work with "white lists" or "black lists" of edges, so that you can for example pre-specify that there is a causal relationship between your exposure and outcome variables. If you are a team of 2, perform a causal inference analysis on both the hand-constructed and the learned DAG, and compare the results. If you are a team of 3, perform an additional causal inference analysis on both the hand-constructed and the learned DAG, and compare the results.
- In each case, you can submit your assignment **once** before the official deadline to get feedback that you can incorporate into your final version. The **deadline** to be guaranteed of feedback is **December 20th, 2024.** You will then receive your feedback at the latest on January 6th, 2025.

For your causal inference analysis, choose from the following list of options:

- Covariate Adjustment
- Instrumental Variable
- Propensity Score Matching or Adjustment
- Case-Control Matching

Note that not every analysis might be applicable to your chosen focal relationship given your DAG. For example there may not always be an instrumental variable that you can use to define your causal effect. Also, to make this assignment interesting, you should choose a focal relationship where there is some confounding, i.e., there should be a back-door path (or several) from exposure to outcome. Choosing such a focal relationship will for example allow you to allow simple (unadjusted) and adjusted or matched analyses of the same relationships.

Report

Write an overall report on your findings. This report should be structured in the usual way of reporting an empirical scientific study, that is, it should contain the following sections:

1. **Abstract:** Summarize your research background, research question, methodology, and results in at most 150 words.
2. **Introduction:** Explain the background of your research in more detail. Explain and cite important prior work. Clearly state your research question. You can re-use content from the first assignment for your introduction, but note that your research question should now consider a "focal relationship": there should be an explicit "exposure" and an explicit "outcome" variable. You can cite your previous assignment 1 as a source for the DAG that you base your causal inference analysis on.
3. **Methods:** Explain the methodologies that you used (2 or 3 chosen from the list above) in enough detail that a competent person (such as a fellow student in the course) would be able to reproduce your results without looking at your code. You don't need to repeat the description of how you constructed your DAG; you can cite the previous assignment with a statement like "We constructed a DAG for this dataset as described previously [...]."
4. **Results:** Show the results of your analyses. Depending on the size of your team, please structure your results in 3 or 4 figures:
   - Figure 1 shows your DAG from assignment 1.
   - Figure 2 shows the results of your first analysis.
   - Figure 3 shows the results of your second analysis.
   - Figure 4 shows the results of your third analysis (for a team of 3).
5. **Discussion:**
   - Summarize your results, and comment on whether they match your expectations. Specifically, explain and comment on the differences between the analyses. If they differ, which ones do you believe the most?
   - If you identify any limitations of the analyses you've performed, can you suggest improvements?

Suggested Algorithms for Structure Learning

If you choose the structure learning option for this assignment: below is a list of structure learning algorithm implementations that you can choose from or use as a starting point for your own research.

- The R package `bnlearn`; particularly, the functions `hc`, `tabu`, and `si.hiton.pc`.
- The R package `pcalg`; particularly, the functions `pc` and `rfci`.
- The R package `lavaan`. While this package does not contain a structure learning algorithm, you can easily use it to create your own by building on the function `modindices`.
- The Python library `pgmpy` (`http://pgmpy.org`), see the function `est.estimate_skeleton` and the function `est.skeleton_to_pdag`.
- The software "GOBNILP", see `https://www.cs.york.ac.uk/aig/sw/gobnilp/`. Apparently a Python wrapper is available at `https://github.com/ncullen93/pyBN`.
- The TETRAD project at `http://www.phil.cmu.edu/projects/tetrad/`, a Java software that implements various algorithms.
- The Bayes net toolbox for Matlab (`https://github.com/bayesnet/bnt`) contains an implementation of the PC algorithm.