Discussion Assignment 2

Katherine Wolf
Introduction to Causal Inference (PH252D)
March 7, 2020

1 Instructions

2 Background

3 Questions to be answered

1. **Specify your observed data.**

   (a) What notation do we use to refer to the distribution of the observed data?
   $\mathbb{P}_{\boldsymbol{O}}$, as in $\boldsymbol{O} = (\boldsymbol{W}, \boldsymbol{A}, \boldsymbol{X}) \sim \mathbb{P}_{\boldsymbol{O}}$.

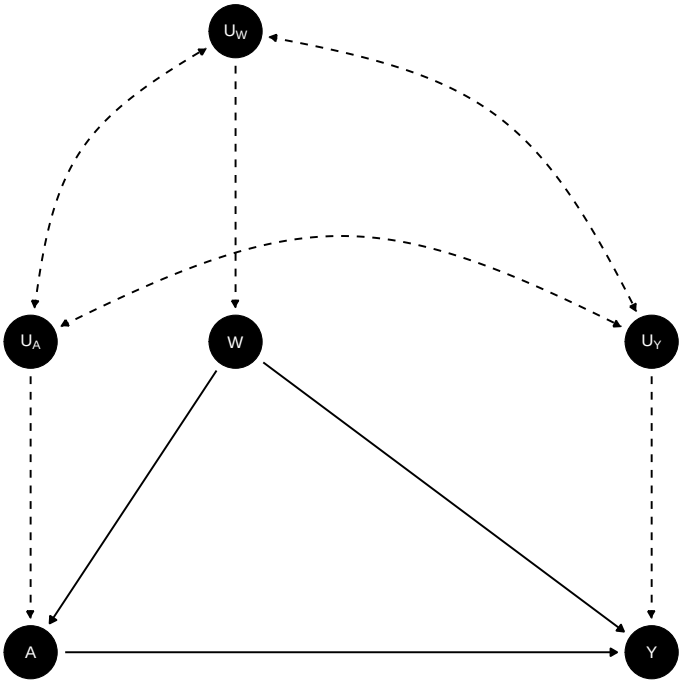
   (b) Specify the link between the SCM and the observed data.
   We link the the observed data to the SCM by assuming that we obtained them from a data-generating system described by our SCM, i.e., that each observation in the data represents a draw from the unknown probability distribution $\mathbb{P}_{\boldsymbol{U}}$ of the exogenous variables $\boldsymbol{U}$, i.e., we drew $\boldsymbol{U} = \boldsymbol{u}$, that we then plugged into our structural equations $\mathcal{F}$ to output a specific $\boldsymbol{X} = \boldsymbol{x}$, of which we observed the subset $\boldsymbol{O} = \boldsymbol{o}$.
   Thus we assume that the the structural equations $\mathcal{F}$ as applied to $\boldsymbol{U} \sim \mathbb{P}_{\boldsymbol{U}}$ will identify the distribution $\boldsymbol{X} \sim \mathbb{P}_{\boldsymbol{X}}$ and, since $\boldsymbol{O} \subseteq \boldsymbol{X}$, the distribution of its observed subset $\boldsymbol{O} \sim \mathbb{P}_{\boldsymbol{O}}$. (Since here we assume that $\boldsymbol{X} = \boldsymbol{O}$, we can write $P_O(O = o) = \sum_u P_f(X = x|U = u)P(U = u)) = \sum_u I(X(u) = x|U = u)P(U = u)$.)

   (c) What is the statistical model $\mathcal{M}$?
   The statistical model $\mathcal{M}$ consists of

   - The endogenous variables $\boldsymbol{X} = (W, A, Y)$;
   - The exogenous variables $\boldsymbol{U} = (U_W, U_A, U_Y) \sim \mathbb{P}_U$; and
   - The structural equations $\mathcal{F}$:

   $$\begin{aligned} W &= f_W(U_W) \\ A &= f_A(W, U_A) \\ Y &= f_Y(W, A, U_Y). \end{aligned}$$

   The structural equations in directed acyclic graph (DAG) format:



Does the SCM place any restrictions on $\mathcal{M}$?

There are no independence assumptions, i.e., no restrictions on the distribution of the exogenous variables $\mathbb{P}_{\boldsymbol{U}}$. Nor are there any exclusion restrictions aside from the ordering of the recursive, time-ordered SCM.

2. **Using the backdoor criterion, assess identifiability of $\Psi^F(P_{\boldsymbol{U},\boldsymbol{X}})$.**

The back-door criterion states that a set of variables $\boldsymbol{Z}$ satisfies it with respect to the variable for the exposure of interest A and the variable for the outcome of interest Y if $\boldsymbol{Z}$ blocks all unblocked back-door (i.e., with an arrow going into A) paths from A to Y without creating any new non-causal associations between A and Y.

Another way to state it is that $\boldsymbol{Z}$ satisfies the back-door criterion with respect to (A, Y) if

- $\boldsymbol{Z}$ blocks all paths from A to Y with an arrow into A, and
- No node in $\boldsymbol{Z}$ is a descendant of A.

If such a set $\boldsymbol{Z}$ exists, then the back-door criterion holds, and thus we can identify the effect of A on Y, specifically the target causal parameter $\Psi^F(P_{\boldsymbol{U},\boldsymbol{X}})$, as a parameter of the observed data distribution $\Psi(\mathbb{P}_O)$ by the G-computation formula.

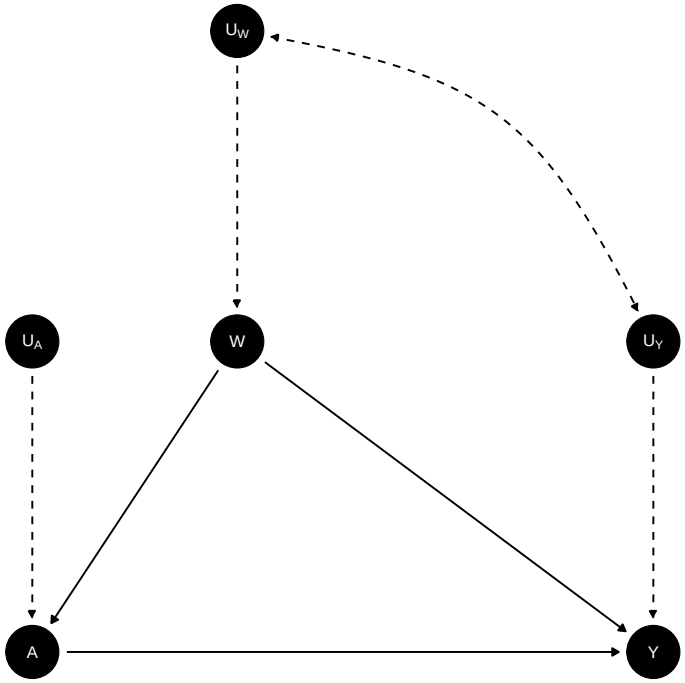
Unfortunately, in the SCM $\mathcal{M}^F$, $\Psi^F(P_{\boldsymbol{U},\boldsymbol{X}})$ is not identified, as the only measured variable available for satisfaction of the back-door criterion is $\boldsymbol{W}$ (or the empty set), and neither $\boldsymbol{W}$ nor the empty set meets the back-door criterion due to the remaining open back-door path between A and Y through U_A and U_Y (including the one through U_A, U_W, and U_Y.)

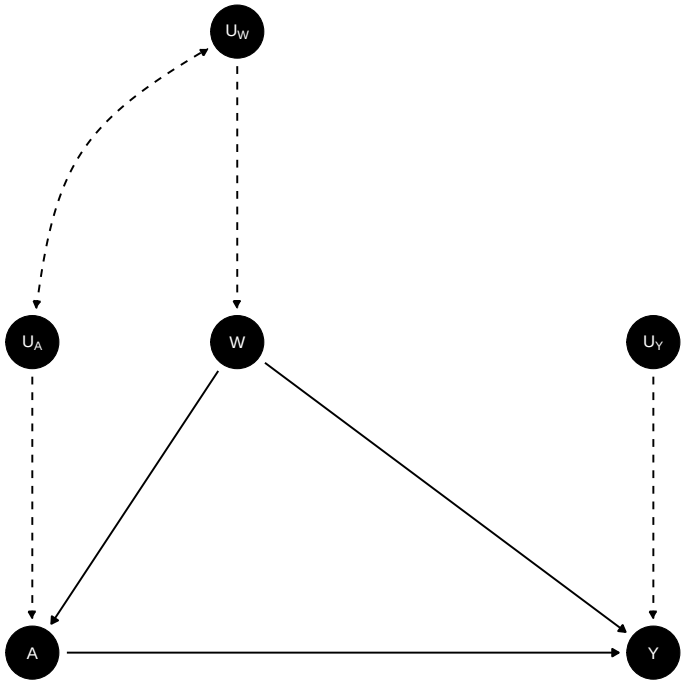
(a) If not identified, under what assumptions would it be? Are some of these sets of additional assumptions more plausible than others? Are there additional measurements you could make so that the needed identifiability assumptions are more plausible?

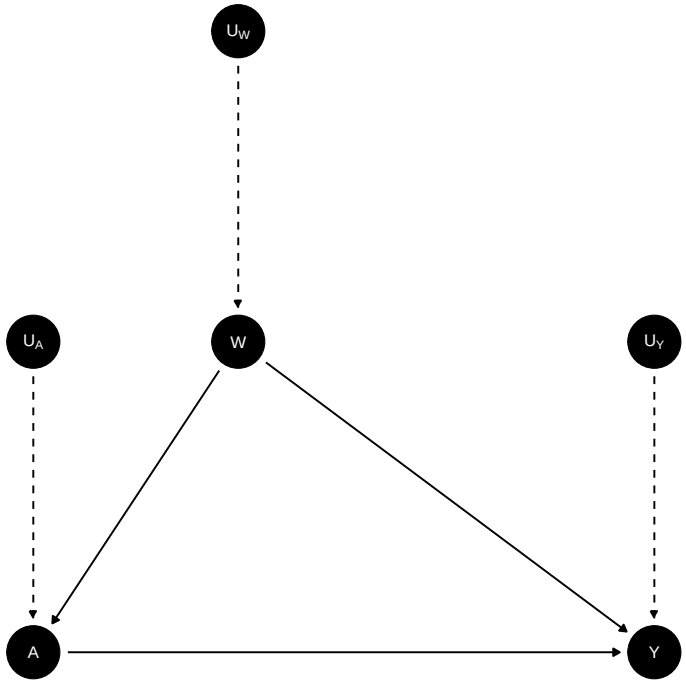
The following DAGs show the three possible working structural causal models that identify $\Psi^F(P_{\boldsymbol{U},\boldsymbol{X}})$.

- $\mathcal{M}_1^{F^*}$ assumptions:
  - U_A is indepdendent of U_Y, i.e., the unmeasured factors U_A that influence energy expenditure A do not affect and are not affected by the unmeasured factors U_Y influencing survival Y

– U_A is indepdendent of U_W, i.e., the unmeasured factors U_A that influence energy expenditure A do not affect and are not affected by the unmeasured factors U_W influencing the covariates like smoking, comorbidities, and body fat W







(b) What notation do we use to denote the original SCM, augmented with additional assumptions needed for identifiability?

$\mathcal{M}^{F^*}$

3. **Specify the target parameter of the observed data distribution (i.e., the statistical estimand).**

4. **What is the relevant positivity assumption? Are you concerned about violations of the positivity assumption in your study?**

   The relevant positivity assumption is

4 Study-specific questions

The investigators assume no unmeasured common causes of (W, A, Y). Is this necessary? Is this sufficient?