Discussion Assignment 2

Katherine Wolf
Introduction to Causal Inference (PH252D)
March 10, 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 $\boldsymbol{O} = (\boldsymbol{W}, \boldsymbol{A}, \boldsymbol{Y}) \sim \mathbb{P}_{\boldsymbol{O}}$.

   (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 the observed data $\boldsymbol{O}$ from some data-generating system compatible with our SCM $\mathcal{M}^{\mathcal{F}}$, i.e., that each observation in the data represents a draw $\boldsymbol{U} = \boldsymbol{u}$ from the unknown probability distribution $\mathbb{P}_{\boldsymbol{U}}$ of the exogenous variables $\boldsymbol{U}$, which 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 applying the structural equations $\mathcal{F}$ 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 $P_{\boldsymbol{O}}(\boldsymbol{O} = \boldsymbol{o}) = \sum_{\boldsymbol{u}} P_f(\boldsymbol{X} = \boldsymbol{x} | \boldsymbol{U} = \boldsymbol{u}) P(\boldsymbol{U} = \boldsymbol{u}) = \sum_{\boldsymbol{u}} I(\boldsymbol{X}(\boldsymbol{u}) = \boldsymbol{x} | \boldsymbol{U} = \boldsymbol{u}) P(\boldsymbol{U} = \boldsymbol{u})$.)

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

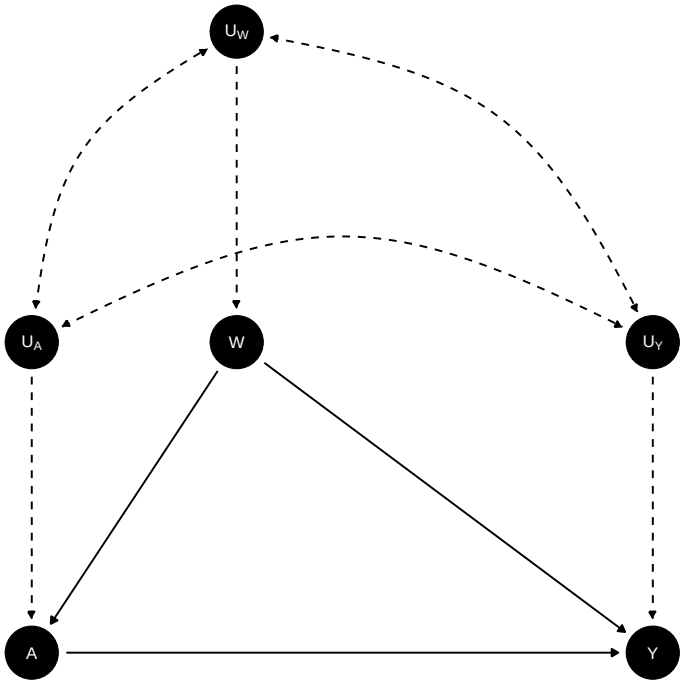
   - The endogenous variables $\boldsymbol{X} = (\boldsymbol{W}, A, Y)$, where
     - $\boldsymbol{W}$ represents the covariates, including (per the first discussion assignment)
       * Smoking
       * Body fat
       * Medical conditions (self-reported and confirmed by treatment and/or medication) including
         · Hypertension
         · Coronary heart disease
         · Myocardial infarction
         · Stroke
         · Lung disease
         · Diabetes
         · Hip or knee osteoarthritis
         · Osteoporosis
         · Cancer
         · Depression;
     - A represents the exposure, energy expenditure, here categorized as binary, where $A = 1$ represents high energy expenditure and $A = 0$ represents low energy expenditure; and

  - Y represents the outcome of seven-year survival;
- The exogenous variables $\boldsymbol{U} = (\boldsymbol{U_W}, \boldsymbol{U_A}, \boldsymbol{U_Y}) \sim \mathbb{P}_{\boldsymbol{U}}$, which represent the unmeasured causes of $\boldsymbol{W}$, A, and Y, respectively; and
- The structural equations $\mathcal{F}$:

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

The structural equations in directed acyclic graph (DAG) format (Figure 1):

Figure 1: Structural causal model (SCM) $\mathcal{M}^F$ represented 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 back-door criterion, assess identifiability of $\Psi^F(P_{\boldsymbol{U},\boldsymbol{X}})$.

The back-door criterion states that for a set of covariates represented by the random variable $\boldsymbol{W}$, an exposure of interest represented by the random variable A, and an outcome of interest represented by the random variable Y, $\boldsymbol{W}$ satisfies the back-door criterion with respect to A and Y if $\boldsymbol{W}$ blocks any association between A and Y that arises from unmeasured common causes, does not create any new non-causal associations between A and Y. and does not block any of the effect of A on Y.

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

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

If such a set $\boldsymbol{W}$ exists, then it meets the back-door criterion, which implies that the randomization assumption $Y_a \perp\!\!\!\perp A|\boldsymbol{W}$ 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 $\Psi(\mathbb{P}_{\boldsymbol{O}})$, by the G-computation formula:

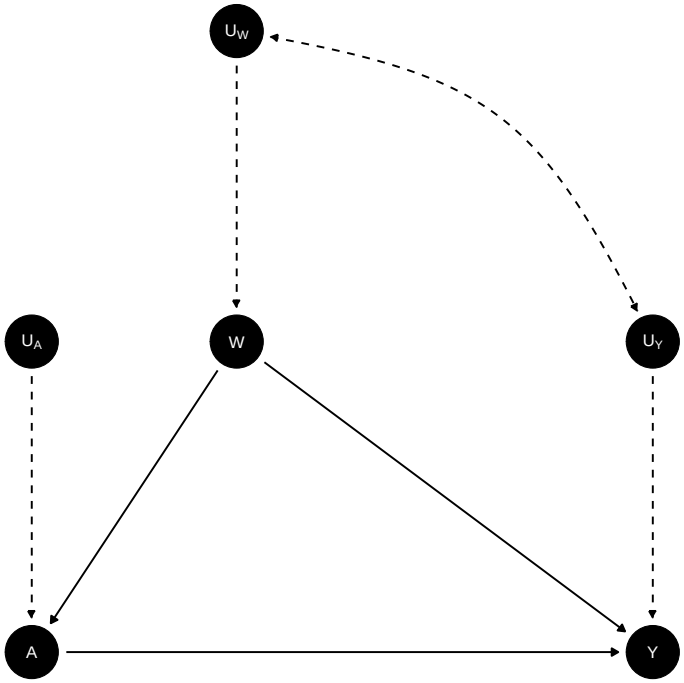
$$\Psi^F(P_{\boldsymbol{U},\boldsymbol{X}}) = E_{\boldsymbol{U},\boldsymbol{X}}(Y_a) = \sum_w E_{\boldsymbol{O}}(Y|A = a, \boldsymbol{W} = \boldsymbol{w}) P_{\boldsymbol{O}}(\boldsymbol{W} = \boldsymbol{w}) = \Psi(\mathbb{P}_{\boldsymbol{O}})$$

Unfortunately, in the SCM $\mathcal{M}^F$, $\Psi^F(P_{\boldsymbol{U},\boldsymbol{X}})$ is not identified as a parameter of $\Psi(\mathbb{P}_{\boldsymbol{O}})$, 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 $\boldsymbol{U_A}$ and $\boldsymbol{U_Y}$ (including the one through $\boldsymbol{U_A}$, $\boldsymbol{U_W}$, and $\boldsymbol{U_Y}$.)

(a) If not identified, under what assumptions would it be?

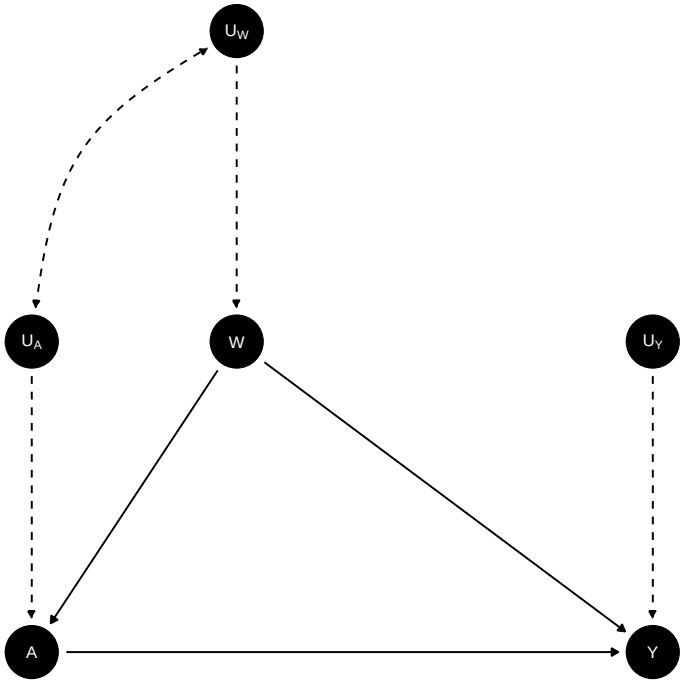
The following DAGs show three possible working SCMs, $\mathcal{M}_1^{F^*}$, $\mathcal{M}_2^{F^*}$, and $\mathcal{M}_3^{F^*}$, that identify $\Psi^F(P_{\boldsymbol{U},\boldsymbol{X}})$, since $\boldsymbol{W}$ satisfies the back-door criterion as described above for all three. (All three assume that $\boldsymbol{U_A} \perp\!\!\!\perp \boldsymbol{U_Y}$, since if $\boldsymbol{U_A}$ and $\boldsymbol{U_Y}$ are associated, no means exists to block that path and thus satisfy the back-door criterion.)

Figure 2: Working structural causal model (SCM) $\mathcal{M}_1^{F^*}$ represented in directed acyclic graph (DAG) format.



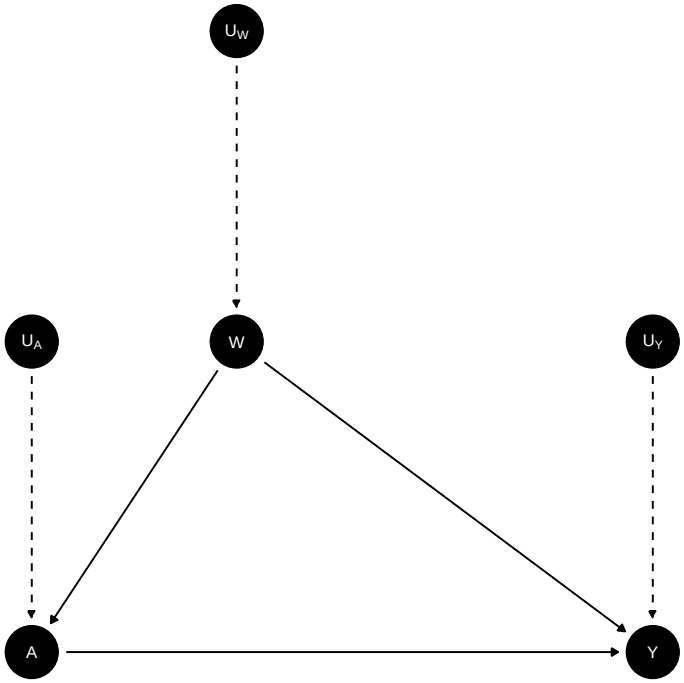
- $\mathcal{M}_1^{F^*}$ assumptions (see Figure 2):
  - $\boldsymbol{U_A} \perp\!\!\!\perp \boldsymbol{U_Y}$, i.e., the unmeasured factors $\boldsymbol{U_A}$ that influence energy expenditure A do not affect and are not affected by the unmeasured factors $\boldsymbol{U_Y}$ influencing seven-year survival Y.
  - $\boldsymbol{U_A} \perp\!\!\!\perp \boldsymbol{U_W}$, i.e., the unmeasured factors $\boldsymbol{U_A}$ that influence energy expenditure A do not affect and are not affected by the unmeasured factors $\boldsymbol{U_W}$ influencing the smoking, body fat, and comorbidity covariates in $\boldsymbol{W}$.

Figure 3: Working structural causal model (SCM) $\mathcal{M}_2^{F^*}$ represented in directed acyclic graph (DAG) format.



- $\mathcal{M}_2^{F^*}$ assumptions (see Figure 3):
  - $\boldsymbol{U_A} \perp\!\!\!\perp \boldsymbol{U_Y}$ (as above).
  - $\boldsymbol{U_W} \perp\!\!\!\perp \boldsymbol{U_Y}$, i.e., the unmeasured factors $\boldsymbol{U_W}$ that influence the smoking, body fat, and comorbidity covariates in $\boldsymbol{W}$ do not affect and are not affected by the unmeasured factors $\boldsymbol{U_Y}$ influencing seven-year survival Y.

Figure 4: Working structural causal model (SCM) $\mathcal{M}_3^{F^*}$ represented in directed acyclic graph (DAG) format.



- $\mathcal{M}_3^{F^*}$ assumptions (see Figure 4):
  All unmeasured factors are independent, i.e.,
  - $\boldsymbol{U_A} \perp\!\!\!\perp \boldsymbol{U_Y}$,
  - $\boldsymbol{U_A} \perp\!\!\!\perp \boldsymbol{U_W}$, and
  - $\boldsymbol{U_W} \perp\!\!\!\perp \boldsymbol{U_Y}$.

Are some of these sets of additional assumptions more plausible than others?

The fewer assumptions made, the more plausible the scenario, making working SCMs $\mathcal{M}_1^{F^*}$ and $\mathcal{M}_2^{F^*}$ more plausible than $\mathcal{M}_3^{F^*}$.

All the working SCMs require the assumption that the unmeasured factors influencing energy expenditure be independent of the unmeasured factors influencing seven-year survival, i.e., that $\boldsymbol{U_A}$ be independent of $\boldsymbol{U_Y}$, or, alternately stated, that all factors influencing both A and Y be measured and included in $\boldsymbol{W}$.

This assumption is likely not justified, as lots of unmeasured social and environmental factors likely influence both energy expenditure and survival but were not included in $\boldsymbol{W}$. Some examples:

- Household wealth might also affect both energy expenditure, as a person with more might have more ability to join a gym, less need to work two jobs and more time for recreational exercise, etc., and survival due to access to a healthier diet, access to preventive healthcare, etc.;
- Neighborhood gun violence could affect both energy expenditure, by reducing participant walking in the neighborhood, and all-cause seven-year survival directly, via exposure to bullets;
- Air pollution around a participant's home might make that participant less likely to exercise and also more likely to die from air pollution exposure;
- Social support might influence energy expenditure (e.g., by inspiring travel to meet friends) and is well documented as a factor that increases survival;
- Age or effects of aging, even witihin the range of ages in the study from 70 to 79, might affect both energy expenditure and seven-year survival.

That said, if a randomized controlled trial, for example, were to randomly assign participants into groups and then assign one group activities that increase energy expenditure and not the other, and were able to influence participants somehow not to alter their activities otherwise, that might make the assumption that $\boldsymbol{U_A}$ is independent of $\boldsymbol{U_Y}$ more plausible.

The working SCMs then also require at least one of two additional assumptions to achieve identifiability. $\mathcal{M}_1^{F^*}$ assumes that the unmeasured factors influencing the covariates listed above neither influence or are influenced by the unmeasured factors influencing the exposure of energy expenditure, i.e., that $\boldsymbol{U_W}$ is independent of $\boldsymbol{U_A}$.

This assumption is also likely not justified. Innumerable unmeasured factors likely affect both energy expenditure A and the presence of various covariates in $\boldsymbol{W}$. Some examples:

- An anxiety disorder might increase both energy expenditure and the incidence of various diseases included in $\boldsymbol{W}$;
- Household wealth might also affect both energy expenditure, as a person with more might have more ability to join a gym, less need to work two jobs and more time for recreational exercise, etc., and might also have a lower likelihood of the diseases in $\boldsymbol{W}$ due to access to a healthier diet, access to preventive healthcare, etc.;
- Neighborhood gun violence could affect both energy expenditure, by reducing participant walking in the neighborhood, and the incidence of some of the diseases, behaviors, or comorbidities in $\boldsymbol{W}$ as ways to cope with or the effects of chronic stress;
- Air pollution around a participant's home might make that participant less likely to exercise and also more likely to develop some of the comorbidities in $\boldsymbol{W}$;
- Social support might influence energy expenditure (e.g., by inspiring travel to meet friends) and the incidence of some of the comorbidities in $\boldsymbol{W}$;
- Age or effects of aging, even given the limitation of the study to participants ages 70 to 79, might affect both energy expenditure and the incidence of some of the comorbidities in $\boldsymbol{W}$.

$\mathcal{M}_2^{F^*}$ assumes that the unmeasured factors influencing the covariates listed above neither influence or are influenced by the unmeasured factors influencing the outcome of seven-year survival, i.e., that $\boldsymbol{U_W}$ is independent of $\boldsymbol{U_Y}$.

This assumption is also likely not justified. Various factors likely affect both the values of the covariates in $\boldsymbol{W}$ and seven-year survival. Some examples:

- A genetic mutation affecting cell repair might both affect cancer risk and might also affect all-cause seven-year mortality directly;

- Household wealth might also affect both the incidence of the comorbidities and smoking behavior in $\boldsymbol{W}$ as well as survival;
- Neighborhood gun violence could affect both the incidence of some of the diseases, behaviors, or comorbidities in $\boldsymbol{W}$ as ways to cope with or the effects of chronic stress as well as survival directly via exposure to bullets;
- Air pollution around a participant's home might affect both the incidence of the comorbidities in $\boldsymbol{W}$ and survival via some other pathway not included in $\boldsymbol{W}$;
- Social support might influence both the incidence of some of the comorbidities in $\boldsymbol{W}$ and survival;
- Age or effects of aging, even given the limitation of the study to participants ages 70 to 79, might influence both the incidence of some of the comorbidities in $\boldsymbol{W}$ and survival.

Are there additional measurements you could make so that the needed identifiability assumptions are more plausible?

Additional measurements that would make the needed identifiability assumptions more plausible might be those evaluating the factors listed above, for the reasons given above. Factors that investigators could measure or data that they could collect via questionnaire include neighborhood of residence (which could be linked to area-level factors like air pollution exposure, gun violence rates, etc.), age, race/ethnicity, anxiety, perceived social support, household wealth, and (controversially) genetic sequencing.

In practice, I would probably use $\mathcal{M}_1^{F^*}$ as the working SCM (which assumes no association between $\boldsymbol{U_A}$ and $\boldsymbol{U_W}$ and no association between $\boldsymbol{U_A}$ and $\boldsymbol{U_Y}$). I would try to measure the factors listed in the prior paragraph and include them in $\boldsymbol{W}$ to make both of those assumptions more plausible. If nothing else, I would choose $\mathcal{M}_1^{F^*}$ because a truly randomized trial forces the independence of $U_A \perp\!\!\!\perp U_W$ and $U_A \perp\!\!\!\perp U_Y$, so a later researcher seeking to confirm conclusions from this observational study using a randomized trial to somehow assign energy expenditure could use a similar working SCM to confirm or challenging my results.

Moreover, should I gain access to additional data, I could check assumptions that $U_A \perp\!\!\!\perp U_W$ in my study data and possibly move factors from $\boldsymbol{U_W}$ to $\boldsymbol{W}$ much earlier and more easily than I could check assumptions that $U_W \perp\!\!\!\perp U_Y$ in my study data, as I would have access to the exposure data A years before the outcome data Y. By the end of the study it might be too late to measure some of the factors in $\boldsymbol{U_W}$ that in my study turned out to be associated with the outcome Y.

Although I don't find any of these sets of assumptions particularly plausible, I also think that the assumption that $\mathcal{M}_1^{F^*}$ does not require, of no shared common causes between $\boldsymbol{U_W}$ and $\boldsymbol{U_Y}$, is perhaps the least plausible given the limited number of behaviors and comorbidities included in $\boldsymbol{U_W}$.

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

$\mathcal{M}^{F^*}$. Henceforth $\mathcal{M}^{F^*}$ refers to $\mathcal{M}_1^{F^*}$ above, which adds the following two assumptions:

- $\boldsymbol{U_W} \perp\!\!\!\perp \boldsymbol{U_A}$
- $\boldsymbol{U_A} \perp\!\!\!\perp \boldsymbol{U_Y}$

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

Under the working SCM $\mathcal{M}^{F^*}$, the target causal parameter of the observed data distribution, i.e., the statistical estimand, is

$$\begin{aligned}\Psi(\mathbb{P}_{\boldsymbol{O}}) &= \mathbb{E}_{\boldsymbol{O}}[\mathbb{E}_O(Y|A=1,\boldsymbol{W}) - \mathbb{E}_{\boldsymbol{O}}(Y|A=0,\boldsymbol{W})] \\ &= \sum_w [\mathbb{E}_{\boldsymbol{O}}(Y|A=1,\boldsymbol{W}=\boldsymbol{w}) - \mathbb{E}_{\boldsymbol{O}}(Y|A=0,\boldsymbol{W}=\boldsymbol{w})]\mathbb{P}_{\boldsymbol{O}}(\boldsymbol{W}=\boldsymbol{w})\end{aligned}$$

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

The relevant positivity assumption is $\min_{a\in\mathcal{A}} P_O(A=a|\boldsymbol{W}=\boldsymbol{w}) > 0$ for all $\boldsymbol{W}$ for which $P_O(\boldsymbol{W}=\boldsymbol{w}) > 0$, i.e., that there must be a positive probability of each exposure level of A within each possible stratum of covariate combinations in $\boldsymbol{W}$. Given the sheer number (12) of factors represented by $\boldsymbol{W}$, even without measuring the additional factors that I think are necessary to include in $\boldsymbol{W}$, and even if each factor in the existing $\boldsymbol{W}$ is binary, that leaves 2^{12} or 4,096

different possible combinations of covariate values in $\boldsymbol{W}$. Thus the assumption that each combination of covariate values actually present in $\boldsymbol{W}$ then exists for at least one treated and one untreated participant seems highly likely to be violated. Thus I am very concerned.

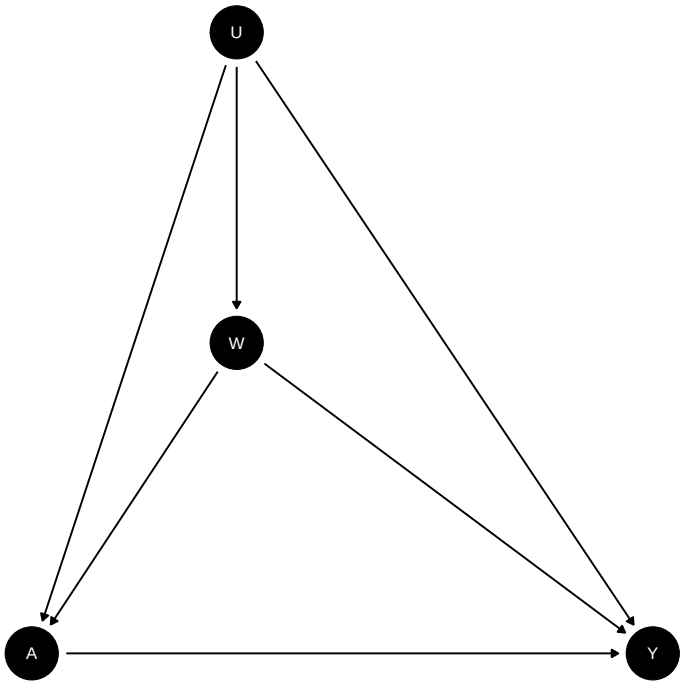
4 Study-specific questions

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

I'm not exactly sure if this is saying that the investigators made the weaker assumption that no single unmeasured cause caused all three, or that the investigators made the stronger assumption that none of the three share any unmeasured cause with another of the three, so I'll answer for both possibilities:

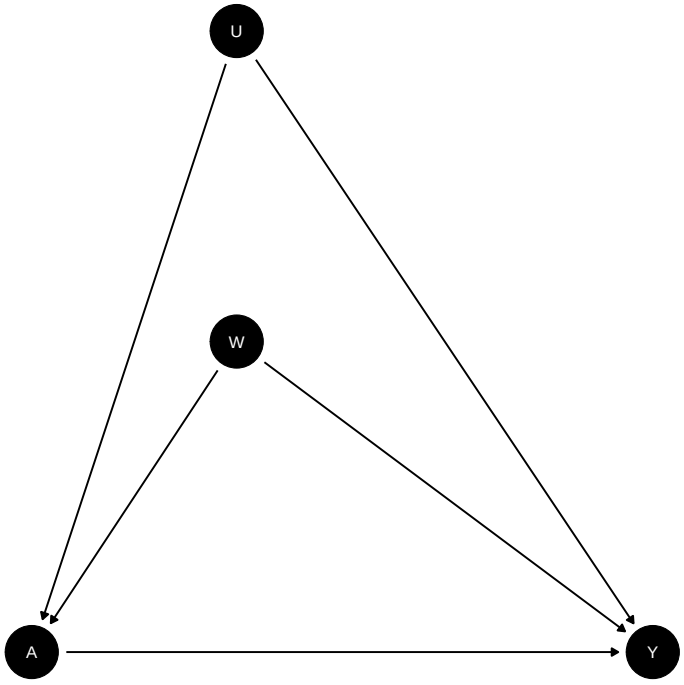
The weaker assumption that no single unmeasured cause exists that causes all three factors represented in the variable set $(\boldsymbol{W}, A, Y)$ is necessary but not sufficient. The case in which such an unmeasured cause does exist that affects all three is a subset of the scenario represented in the structural causal model of Figure 1, where all the unmeasured causes are associated with all other unmeasured causes. (Those associations could represent identity, i.e., that a single random variable representing the same cause is an element in all three random variables $\boldsymbol{U_W}$, $\boldsymbol{U_A}$, and $\boldsymbol{U_Y}$). The minimal situation, in which only one unmeasured cause U exists but is a cause of $\boldsymbol{W}$, A, and Y, we can represent as in Figure 5.

Figure 5: Structural causal model where only one unmeasured cause U exists but causes all three of $\boldsymbol{W}$, A, and Y.



Because U is unmeasured, we cannot block the path through U from A to Y by controlling for it, so we cannot satisfy the back-door criterion. Thus we cannot use the G-computation formula to identify the effect of energy expenditure A on all-cause seven-year survival Y with the data we have measured. We would need to measure the cause U affecting all three. Thus, the assumption that no such U exist is necessary. That said, though, the nonexistence of such a U is not sufficient. Take the hypothetical case portrayed in Figure 6, where a common cause of A and Y exists that is not a cause of $\boldsymbol{W}$:

Figure 6: Structural causal model where only one unmeasured cause U exists and causes A and Y but not $\boldsymbol{W}$.



Because U is unmeasured, we cannot block the path through U from A to Y by controlling for it, so we cannot satisfy the back-door criterion. Thus we cannot use the G-computation formula to identify the effect of energy expenditure A on all-cause seven-year survival Y with the data we have measured. We would need to measure the cause U affecting both A and Y. Thus, the assumption that no U exists that affects all three of $\boldsymbol{W}$, A, and Y is not sufficient to identify the target casual parameter as a parameter of the observed data distribution. We've shown by counterexample a situation in which no such U exists but in which we still cannot satisfy the back-door criterion or use the G-computation formula to identify the effect of energy expenditure A on all-cause seven-year survival Y with the data we have measured.

The stronger assumption that none of the three factors represented in the variable set $(\boldsymbol{W}, A, Y)$ shares a common cause with another variable in the set is sufficient but not necessary. This case is represented by $\mathcal{M}_3^{F^*}$ in Figure 4. In this case, $\boldsymbol{W}$ blocks all paths with an arrow into A without including any descendants of A, so it satisfies the back-door criterion, and we can use the G-computation formula to identify the target causal parameter $\Psi^F(P_{U,\boldsymbol{X}})$ as a parameter of the observed data distribution $\Psi(\mathbb{P}_{\boldsymbol{O}})$. Thus the assumption is sufficient.
The working SCMs $\mathcal{M}_1^{F^*}$ and $\mathcal{M}_2^{F^*}$ in figures 2 and 3, however, show that this assumption is not necessary, as the target causal parameter $\Psi^F(P_{U,\boldsymbol{X}})$ is also identifiable as a parameter of the observed data distribution $\Psi(\mathbb{P}_{\boldsymbol{O}})$ under those working SCMs despite the potential existence of a common cause of $\boldsymbol{W}$ and Y in $\mathcal{M}_1^{F^*}$ (Figure 2) and the potential existence of a common cause of $\boldsymbol{W}$ and A in $\mathcal{M}_2^{F^*}$ (Figure 3).

5 Questions for GSIs

Hi, GSIs!
Maya's new slides for Lecture 5 appear to have a blank slide at the end labeled "Positivity for Continuous W" but no actual content on that slide. :(