Project Presentation

Alex Luedtke, Lucia Petito, Steven Pollack

PHC252D

Outline

- Background
- Specify SCM (and DAG)
- Specify counterfactuals and target causal quantity
- Introduce data and commit to a statistical model
- Discuss identifiability and estimand
- Get our hands dirty (estimation procedures)
- Results
- Interpretation

Background

We know that sleep affects weight, but does trying to lose weight affect sleep?

- We used National Health and Nutrition Examination Survey (NHANES) data – from the National Center for Health Statistics (NCHS) – a multistage survey of U.S. population
  - Stage 1: Counties
  - Stage 2: Segments
  - Stage 3: Households
  - Stage 4: Individuals
- Survey aims to study wide range of topics such as Cardiovascular disease, Obesity, Physical fitness and physical functioning, Reproductive history and sexual behavior, etc.

Background (Cont'd)

Notes about NHANES data:

- Individuals were subjected to interviews as well as physical examinations.
  - categorical as well as numerical data
  - some questions had a lot of valid responses, but made positivity questionable.
  - No shortage of missing data (either "I don't know"'s or unanswered questions).
- The sample for the survey is selected to represent the U.S. population of all ages. To produce reliable statistics, NHANES over-samples persons 60 and older, African Americans, and Hispanics.

W: Baseline Covariates

- Gender
- Age in months (300-959 months, 25-79 years)
- Race/Ethnicity (Mexican American, Other Hispanic, Non-Hispanic White, Non-Hispanic Black, Other)
- Education Level (less than high school, high school/GED, some college, college and above)
- Marital Status (never married, married/living with partner, divorced/separated)
- Annual Household Income (less than or greater than $20k)
- Body Mass Index (continuous from 15-50)

A: Exposure Variable

- The subject's response to the question: "During the past 12 months, have you tried to lose weight?"
- Note that this does not restrict to dieting

Y: Response Variable

- The subject's response to the question: "How much sleep do you usually get at night on weekdays or workdays?
- Both A and Y sampled simultaneously, so temporal ordering is only assumed

SCM and DAG

Our observational data structure is $O = (W, A, Y) \sim P_0$. With mild temporal assumptions, one SCM is:

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

Note: no assumptions made on functional forms of W, A, or Y.

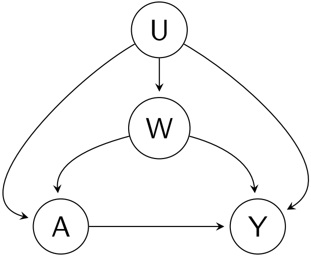


Figure: Simplified DAG – no independence assumptions on U's.

W: Baseline Covariates

- Gender
- Age in months (300-959 months, 25-79 years)
- Race/Ethnicity (Mexican American, Other Hispanic, Non-Hispanic White, Non-Hispanic Black, Other)
- Education Level (less than high school, high school/GED, some college, college and above)
- Marital Status (never married, married/living with partner, divorced/separated)
- Annual Household Income (less than or greater than $20k)
- Body Mass Index (continuous from 15-50) – as of the prior year

Counterfactuals and Causal Quantities...

- Since the intervention is a point treatment, our counterfactual is Y_a: the average sleep one would get, if they had (or had not) attempted to lose weight.

- We are interested in measuring the ATE of attempted weight lose on average sleep:

$$\Psi(P_{U,X}) = \mathbb{E}_{U,X}[Y_1] - \mathbb{E}_{U,X}[Y_0]$$

Statistical Models

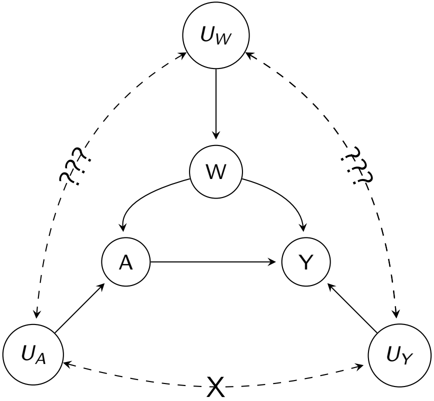
- No assumptions on the functional forms of f_U, f_A, f_Y.
- No, a priori established, assumptions on independence between any of the U's:
  - It's plausible that household income is a mediator for the effect of workplace stress on sleep.
- Choose the (non-parametric) model, $\mathcal{M}$, of distributions compatible with our SCM.

Identifiability

- Since Ψ is the ATE, we need to satisfy backdoor criterion to identify g-computation with Ψ:

$$Y_a \perp\!\!\!\perp A \mid W$$

- Need $U_A \perp\!\!\!\perp U_Y$ and either
  - $U_A \perp\!\!\!\perp U_W$ (semi-plausible)
  - $U_W \perp\!\!\!\perp U_Y$ (less plausible)



Estimands

Provided we can accept $U_A \perp\!\!\!\perp U_W$ or $U_W \perp\!\!\!\perp U_Y$, we have:

- Simple Substitution:

$$\psi_0 \approx \frac{1}{n}\sum_{i=1}^{n} \widehat{\mathbb{E}}\left[Y \mid A = 1, W = W_i\right] - \widehat{\mathbb{E}}\left[Y \mid A = 0, W = W_i\right]$$

- IPTW:

$$\psi_0 \approx \frac{1}{n}\sum_{i=1}^{n} \left(\frac{I(A_i = 1)}{\hat{g}_n(A_i \mid W_i)} - \frac{I(A_i = 0)}{\hat{g}_n(A_i \mid W_i)} \right) Y_i$$

- TMLE:

$$\psi_0 \approx \frac{1}{n}\sum_{i=1}^{n} \left(\bar{Q}_n^1(1, W_i) - \bar{Q}_n^1(0, W_i)\right)$$

Estimation Procedures

Estimate $\bar{Q}_n^0(a, w) = \widehat{\mathbb{E}}(Y \mid A = a, W = w)$ (and $\hat{g}_n(a \mid w)$) via Super Learner with library:

SL.mean	$Y \sim A$
SL.earth	$Y \sim A \times Gender \times RaceEth + MarStat \times HHInc$
SL.rpartPrune	$+AgeMonths{:}Gender + EduLevel + AgeMonths$
SL.ridge	$Y \sim A \times Gender \times MarStat$
SL.glmnet	$Y \sim A \times Gender \times AgeMonths \times HHInc$
median	$Y \sim A \times ExamDate \times MarStat$

Table: Super Learner library

Is the : a typo? And we'll estimate the treatment mechanism using the same library and glm's but without regressing on A.