The effect of "attempting to lose weight" on sleep

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 habits affect weight, but does trying to lose weight affect sleep patterns?

- Doctors advise consistently getting enough sleep each night to maintain a healthy lifestyle.
- Previous research has shown an association between sleep deprivation and obesity and oversleeping and obesity (indicates a U-shaped curve).
- Hormonal changes from dieting could cause disruptions in sleep cycle (could also be caused by exercise).

- 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

Picture courtesy of NHANES

- 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 practical 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, adolescents, 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 from one year ago (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

Massaging the Data

- Deleted all people who refused to identify Race/Ethnicity, Education Level, Marital Status, Annual Household Income, dieting status

- Forced to coarsen the categories in Annual Household Income

SCM and DAG

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

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

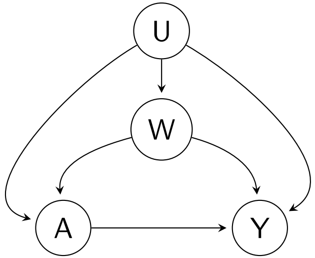


Figure: Simplified DAG – no independence assumptions on U's.

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

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

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

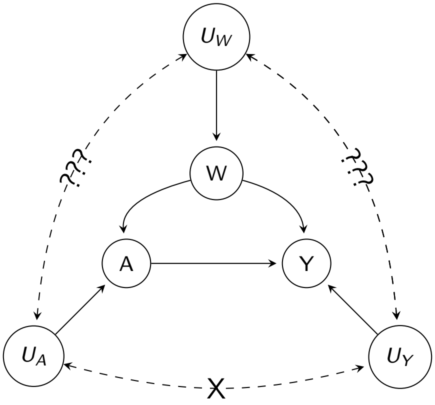
$$\Psi(P_{U,X}) = \mathbb{E}_{U,X}[Y_1] - \mathbb{E}_{U,X}[Y_0]$$

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)



Statistical Models

- No assumptions on the functional forms of f_U, f_A, f_Y.
- Choose the (non-parametric) model, $\mathcal{M}$, of distributions compatible with our SCM.

Positivity Concerns

- We do not believe that trying to lose weight or not trying to lose weight occurs with probability 0 given any set of covariates W
- Practical positivity violations are a concern given W is high dimensional

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	$Y \sim A \times Gender \times AgeMonths \times HHInc$

Table: Super Learner library

Preliminary Results

- Multistage survery designed makes bootstraping more involved than usually the case.
- Individual's have survey weights associated to them, however we're unsure how to incorporate these weights during analysis
- High dimensionality of W makes assessing positivity difficult.
- The following analysis was performed using "simple random sampling" procedure during the bootstrap and equal weighting for individuals.

Results (Cont'd)

- Practical positivity violations for now only assessed on categorical variables $V \subset W$
  - Gender, Race/ethnicity, Education Level, Marital Status
- No members in our sample who tried to lose weight given:

Gender	Race	Education	Marital Status
Male	Other	Less than HS	Unmarried/not living with partner
Female	Other	HS	Married/living with partner

Results (Cont'd)

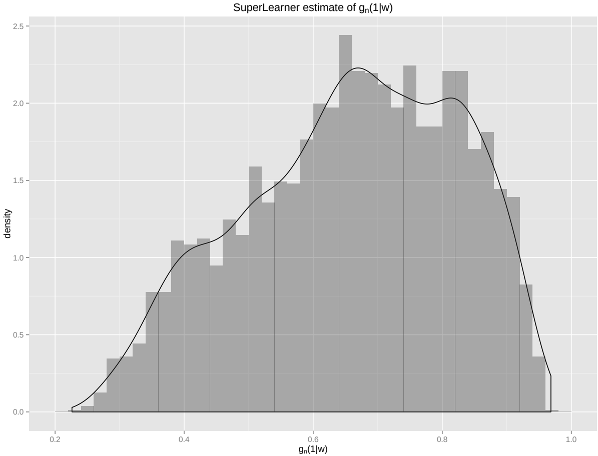


Figure: Histogram and Density estimate from SuperLearner output of g_n.

Results (Cont'd)

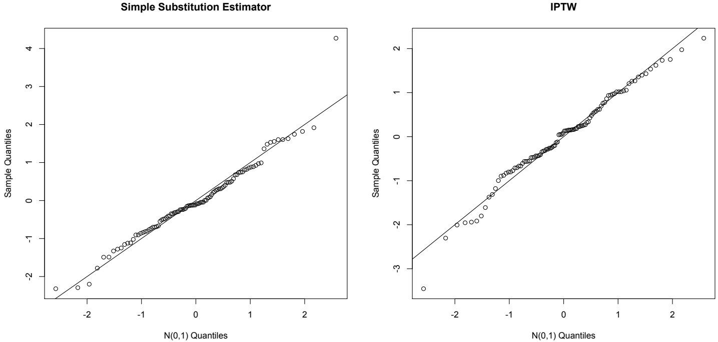
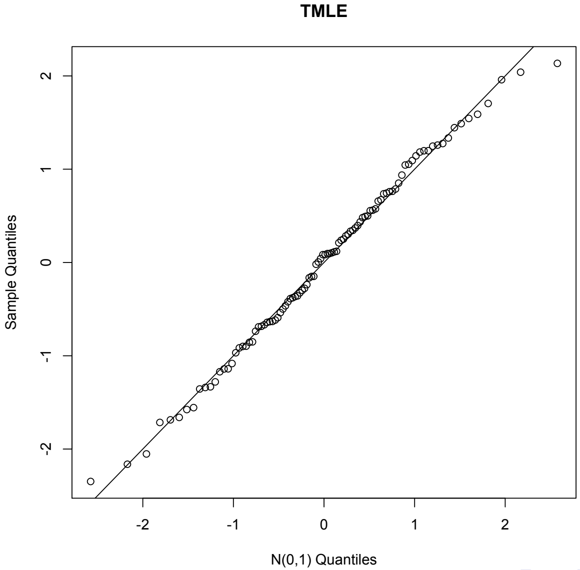


Figure: Q-Q plots for G-Comp and IPTW estimators against Normal

Results



Results (Cont'd)

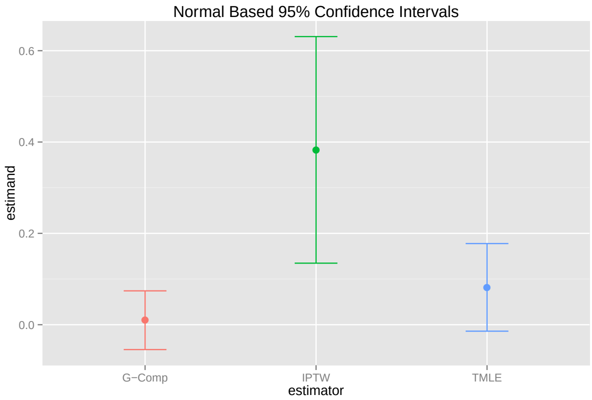


Figure: 95% Confidence intervals based on Normal approximation...

Results (Cont'd)

	Estimate	95% Normal CI	95% Quantile CI
G-Comp	0.00983	(-0.05444, 0.07411)	(-0.04403, 0.07961)
IPTW	0.38282	(0.13473, 0.63091)	(0.40209, 0.89568)
TMLE	0.08175	(-0.01418, 0.17769)	(-0.01667, 0.16519)

Table: Results for naïve small sample bootstrap

Results (Cont'd)

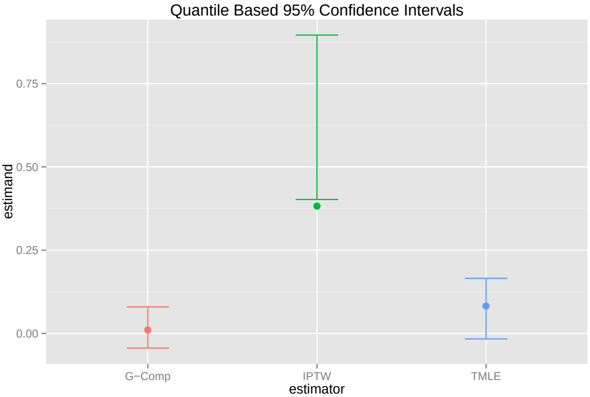


Figure: 95% Confidence intervals based off of Bootstrap Quantiles...

Shortcomings

- Removed all missing data on a complete-case basis: deleted 2,832 people
- Should have more covariates
- Impossible to accurately bootstrap as it stands: did not account for survey weights
- Did not account for survey weights in our GLMs: not yet sure how to
- Difficult temporal ordering
- Identifiability implausible (U_W probably not independent of U_Y)
- Covariates do not have predictive power for Y: SuperLearner performs similarly to basic GLM