Discussion Assignment #2

Sabrina Boyce, Shelley Facente, and Steph Holm

10/30/2019

Question 1: What is the goal of the article?

The goal of the article is to provide an example of using the parametric g-computation formula to estimate the impact of various behavioral and lifestyle interventions on CHD risk using data from the Nurses Health Study.

Question 2: Consider the effect of diet on the cumulative risk of coronary heart disease (CHD) while controlling for a time-dependent diabetes. The causal model is represented by the following DAG. For simplicity, we have omitted the exogenous factors U and ignored censoring as well as death by other causes.

Question 2(a): Why is diabetes a time-dependent confounder?

Diabetes is a time-dependent confounder because:

1. Diabetes is a risk-factor for the outcome, CHD
2. A diagnosis of diabetes could impact future exposure status, diet
3. Prior diet may impact whether or not a person is diagnosed with diabetes.

So, this covariate can both be impacted by the exposure and can impact the exposure.

Question 2(b): Specify the corresponding structural causal model $\mathcal{M}^F$ using the following notation:

- Let L_k be diabetes status at time k.
- Let A_k be diet at time k. Then we can set it according to a static intervention: "good" $A_k = 1$ or "bad" $A_k = 0$ at time k.
- Let Y_k be an indicator that the subject was diagnosed with CHD by time k.

$$O = \{\bar{L}(k), \bar{A}(k), Y(k)\}$$

$$\bar{L}(k) = \text{Diabetes diagnosis at time k}$$

$$\bar{A}(k) = \text{diet at time k}$$

$$Y(k) = \text{CHD at time k}$$

$$\bar{L}(k) = f_{L(k)}(U_{\bar{L}(k)}, \bar{A}(k-1), \bar{L}(k-1))$$

$$\bar{A}(k) = f_{A(k)}(U_{\bar{A}(k)}, \bar{A}(k-1), \bar{L}(k), \bar{Y}(k-1))$$

$$Y(k) = f_{Y(k)}(U_{\bar{Y}(k)}, \bar{A}(k), \bar{L}(k), \bar{Y}(k-1))$$

Question 2(c): Specify the counterfactuals of interest and the target causal parameter.

Our counterfactuals of interest are the values of $Y_{\bar{a}_k}$ we would have observed if A_0 had been set to values 0, 1, 2, 3,...18 at each time point, all the 18 possible lifestyle interventions at each time point. We are able to generate these counterfactuals by plugging in every combination of treatment level across all time points and observing the outcome for each.

The target causal parameter is given by the following marginal structural model:

$$E(Y_{\bar{a}_k}) = m(\bar{a}|\beta) = logit(P_{U,X}(Y_{\bar{a}=1})) = \beta_0 + \beta_1 a$$

Our target causal parameter is β_1 such that e^{β_1} would give us the causal OR for CHD associated with the 18 lifestyle interventions. β_1 represents the causal parameter if the relationship between the interventions and CHD is unconfounded by measured or unmeasured factors.

Question 2(d) What are the observed data O? Factorize the observed data distribution P0 according to the time-ordering.

N i.i.d. copies of $O = (A(\bar{t}), L(\bar{t})) \sim P_0$

The statistical estimand is

$$\Psi(P_0) = \sum_{\bar{l}^1 0}(E(Y(k)|\bar{A} = 1, \bar{L} = \bar{l}, \bar{Y}(k-1) = \bar{y}(k-1))xP(L(k) =$$

$$\sum_{\bar{l}}(E(Y(k)|\bar{A} = 1, \bar{L} = \bar{l}, \bar{Y}(k-1) = \bar{y}(k-1))xP($$

Question 2(e): What assumptions are needed to identify the causal parameter from the observed data distribution?

In order to identify our causal parameter from the observed data distribution, we must use the:

1. Randomization assumption, which assumes a is independent of y.
2. Sequestion backdoor path criteria which allows us to control for a different set of covariates at each time point.
3. Positivity assumption, which assumes that no strata have few or no observations.
4. No unmeasured confounding
5. Correct model is specified
6. No measurement error

Question 2(f): Specify the G-computation formula (i.e. the statistical estimand) for this example.

$$\hat{\psi}_{n,gcomp}(P_n) = 1/n \sum_{i=1}^{n} \bar{Q}_n(A_i = a, W_i)$$

Question 2(g): How would you implement a longitudinal parametric G formula (i.e. non-targeted substitution) estimator for this parameter?

1. Use a pooled logistic regression to model each time-varying risk factors using time-varying and baseline risk factors for the 2 previous time points (2 years and 4 years prior) among those with no prior diagnosis of CHD
2. Use a pooled logistic regression to model CHD using time-varying and baseline risk factors for the 2 previous time points (2 years and 4 years prior) among those with no prior diagnosis of CHD Take the product of the conditional probabilities of each time-varying risk factor
3. Simulate data (n=10,000) using Monte Carlo simulation
4. Simulate covariates by predicting their values using the risk factor models' coefficients from step 1.
5. Use a logistic regression model to model each level of the intervention using time-varying and baseline risk factors.

Question 2(g): How would you implement a longitudinal parametric G formula (i.e. non-targeted substitution) estimator for this parameter?

6. Predict CHD at each time point for each individual based on the simulated and intervened upon risk factors.
7. Calculate the cumulative risk across all individuals
8. Repeat 6 and 7 above for each level of A to get a comparison of A=1 and A=0
9. Use bootstrapping to calculate 95% confidence intervals

Question 3. Discuss the hypothetical interventions on risk factors for CHD.

3(a) How are interventions 1 and 2 different?

3(b) Why do the authors consider the BMI intervention to be “unrealistic”?

Question 4. How is censoring handled in the primary analysis? Discuss the relevant assumptions.

Question 5. How is death by another cause handled in the primary analysis?

Question 6. What exclusion restrictions do the authors make?

Question 7. What parametric assumptions do the authors make? How much might these impact results (alluded to in the sensitivity analyses)? How might you approach the estimation problem differently?

Question 8. On page 1604, the authors describe the simulation of risk factor histories not present in the observed data.What assumption does this refer to? What is the danger here?

Question 9. Bonus: The authors say they used cross validation to fine tune prediction models. What do you think they mean? How would you use cross validation to fine tune prediction models?