Discussion Assignment #2

Sabrina Boyce, Shelley Facente, and Steph Holm

10/23/2019

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 time-dependent diabetes.

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, and
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.

$$
\begin{aligned}
O &= \{\bar{L}(k), \bar{A}(k), Y(k)\} \\
\bar{L}(k) &= \text{Diabetes diagnosis at time k} \\
\bar{A}(k) &= \text{diet at time k} \\
Y(k) &= \text{CHD at time k}
\end{aligned}
$$

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

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, with these values representating 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, where e^{β_1} gives 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 P_0 according to the time-ordering.

N i.i.d. copies of $O = (L_0, A_0, L_1, A_1, L_2) \sim P_0$

(with Y as a subset of the Ls)

Basic probability reminder for factoring:

- $P(A, B) = P(A \cup B) = P(A|B) \times P(B)$
- If we know that say 10% of the class drinks coffee, and that of coffee drinkers 50% like ginger cookies, then 5% of the class drinks coffee AND likes ginger cookies.
- So then, $P(A, B, C) = P(A|B, C) \times P(B|C) \times P(C)$

Because we have something that is a function of the joint distribution (P_0) we can factorize it using probability rules into conditional distributions that we can use.

Question 2(d) What are the observed data O? Factorize the observed data distribution P_0 according to the time-ordering.

$$\begin{aligned}
&P(L_0, A_0, L_1, A_1, L_2) = \\
&P(L_2|A_1 = a_1, L_1 = l_1, A_0 = a_0, L_0 = l_0) \times \\
&P(A_1|L_1 = l_1, A_0 = a_0, L_0 = l_0) \times P(L_1|A_0 = a_0, L_0 = l_0) \times \\
&P(A_0|L_0 = l_0) \times P(L_0 = l_0)
\end{aligned}$$

$$\begin{aligned}
&P(L_0, A_0, L_1, A_1, L_2) = \\
&\quad \prod_0^2 P(L_k = l_k|\bar{A}_{k-1} = \bar{a}_{k-1}, \bar{L}_{k-1} = \bar{l}_{k-1}) \times \\
&\quad \prod_0^1 P(A_k = a_k|\bar{A}_{k-1} = \bar{a}_{k-1}, \bar{L}_{k-1} = \bar{l}_{k-1})
\end{aligned}$$

Question 2(d) What are the observed data O? Factorize the observed data distribution P_0 according to the time-ordering.

$$P(O = o) = \prod_{0}^{2} Q(L_k = l_k | \bar{A}_{k-1} = \bar{a}_{k-1}, \bar{L}_{k-1} = \bar{l}_{k-1}) \times \prod_{0}^{1} g(A_k = a_k | \bar{A}_{k-1} = \bar{a}_{k-1}, \bar{L}_{k-1} = \bar{l}_{k-1})$$

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 assume:

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

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

The general longitudinal G computation formula is:

$$E(Y_{\bar{a}}) = \sum_{\bar{l}} E(Y | \bar{A}_K = \bar{a}_K, \bar{L}_K = \bar{l}_K) \times \prod_{t=1}^{K} P(L_t = l_t | \bar{A}_{t-1} = \bar{a}_{t-1}, \bar{L}_{t-1} = \bar{l}_{t-1})$$

Therefore in this specific example it would be:

$$E(Y_{\bar{a}}) = \sum_{\bar{l}} E(Y | \bar{A}_2 = \bar{a}_2, \bar{L}_2 = \bar{l}_2) \times \prod_{0}^{2} P(L_t = l_t | \bar{A}_{t-1} = \bar{a}_{t-1}, \bar{L}_{t-1} = \bar{l}_{t-1})$$

(Given that timepoints run from $t = 0$ to $t = 2$, and thus $K = 2$)

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 model to model non-CHD death for the 2 previous time points (2 years and 4 years prior) among those with no prior diagnosis of CHD
3. 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
4. Simulate data (n=10,000) using Monte Carlo simulation
5. Simulate covariates by predicting their values using the risk factor models' coefficients from step 1.
6. Use a logistic regression model to model each level of the intervention using time-varying and baseline risk factors.
7. Predict the probability of non-CHD death at each time point based on the simulated and intervened upon risk factors.
8. Predict CHD at each time point for each individual based on the simulated and intervened upon risk factors.
9. Calculate the cumulative risk across all individuals
10. Repeat 4-9 above for each level of A to get a comparison of A=1 and A=0
11. 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?

Intervention 1 is to avoid smoking, wheras intervetion 2 is to exercise at least 30 minutes each day. Obviously these are practically very different interventions, but importantly, intervention 1 is not something that could be intervened upon in real life. Once someone has smoked, they can't ever be a never smoker again.

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

The BMI intervention is specifying "an attribute rather than a behavior" so the method of intervening on this would be more ambiguous, and different interventions could affect CHD differently.

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

For the censored population, the authors calculated the estimated 20-year risk of CHD based on their covariate pattern and used this to estimate the cumulative incidence among those who had been censored.

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

The authors treated death from other causes the same as censoring; they calculated the estimated 20-year risk of CHD based on the covariate pattern of subjects and used this to estimate the cumulative incidence

among those who had died from other causes.

Question 6. What exclusion restrictions do the authors make?

The authors make no exclusion restrictions.

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?

The authors use a parametric model for probability of CHD and risk of death given a variety of time-varying risk factors (Ls): pooled logistic regression. The order in which the covariates are included in the regression model affects the risk estimates. Model misspecification may also bias the estimates because of improper extrapolation beyond the limits of the data, or improper fitting of a line or curve based on assumed relationships of the treatment and outcome.

I might approach this estimation problem using nonparametric estimators, such as TMLE. This would allow us to utilize SuperLearner or other strategies to generate estimates, not subject to model misspecification in the same ways.

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?

This assumes that average risk factor histories of people who are censored (or died from non-CHD causes) are not meaningfully different from the average risk factor histories of people who continue to be observed in the dataset, such that the missing data can be filled in via covariate patterns. The danger here is that there could be meaningful differences between people who are censored or died from other causes, and violations of the assumption used here could obscure important patterns in risk of death as a result of CHD in this population.

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?