Maximum Likelihood Estimation of the Longitudinal G-Comp. Estimand

PH252E - Advanced Topic in Causal Inference

Assigned: October 16, 2019
Present in class: October 23, 2019
Write up Due: October 30, 2019

Please read the below article and come prepared to present the following questions with your group.

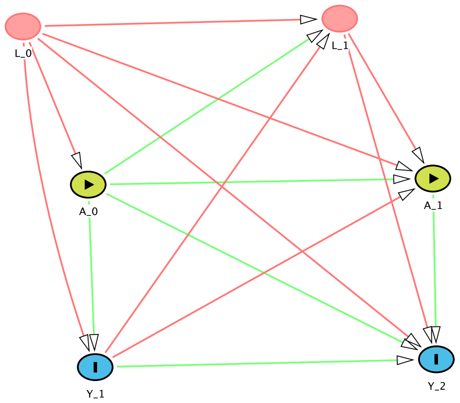
Required Reading:

S.L. Taubman, J.M. Robins, M.A. Mittleman, and M.A. Hernan. Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. Int J Epidemiol, 38(6):15991611, 2009. PMCID: PMC2786249

OK to focus your reading on static, non-responsive interventions. For example, we intervene at all time points to prevent smoking (with probability 1) or to prevent censoring (with probability 1). Later in the semester, we will go over on dynamic (i.e. subject responsive) regimes and (time-permitting) stochastic interventions.

1. What is the goal of the article?

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.



   (a) Why is diabetes a time-dependent confounder?

   (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 to be an indicator that the subject was diagnosed with CHD by time k.

   The authors use slightly different notation. Specifically, the authors define D_k to be an indicator that the subject was diagnosed with CHD in the interval from $(k-1, k]$.

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

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

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

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

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

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

   (a) How are interventions 1 and 2 different?

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

4. How is censoring handled in the primary analysis? Discuss the relevant assumptions.
5. How is death by another cause handled in the primary analysis?
6. What exclusion restrictions do the authors make?
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?
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?
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?