Model-Based Inference for Completely Randomized Experiments

Lorenzo Lagos
February 22, 2017

The main takeaway from this chapter that any causal estimand is, implicitly or explicitly, based on the **imputation** of the unobserved potential outcomes. The model-based approach relies on two primitives: the assignment mechanism and the joint distribution of the potential outcomes. The goal is to calculate the conditional distribution of the missing potential outcomes given the observed data—$f(Y^{mis}|Y^{obs}, W)$—in order to infer the distribution of the estimand of interest $\tau(Y(0), Y(1), W)$. This chapter has a randomized experiment as its running example, so he assignment mechanism plays a minor role. In my project, which is an observational study, the same approach can be taken but the resulting inference will be sensitive to the modeling assumptions.

Assuming the assignment mechanism is random and there are no covariates X, the authors take a **Bayesian approach** to estimate $f(Y^{mis}|Y^{obs}, W)$. For this they require two inputs: a model of the potential outcomes given a vector of parameters—$f(Y(0), Y(1)|\theta)$— and a prior distribution of the parameter vector $p(\theta)$. The former requires specific subject-matter knowledge (e.g., economic theory). The choice of prior in most cases has no substantive influence on the conclusions. In the case of an observational study, the third input is the assignment mechanism $f(W|Y(0), Y(1))$.

The authors outline 4 analytical steps for getting to the distribution of the estimand of interest, but for our purposes the **simulation method** seems more appropriate. By simply using the posterior distribution of the missing data given the observed data and the parameters (Step 1), and then deriving the posterior distribution of the parameters given the observed data (Step 2), we can repeatedly impute the missing potential outcomes. Given advances in computation, this approach may be much simpler than going through the analytic solution.

Two interesting points I got from the chapter. First, that adding the **correlation coefficient** as a parameter in the model will add uncertainty since the data contain no information about the correlation between potential outcomes. Being conservative about this can take the form of assuming $\rho = 1$ (worst case posterior variance) or $\rho = 0$ (avoid contamination of the imputation of potential outcomes). Second, when **adding covariates** to the analysis, the set of parameters for covariates θ_X may be different from those for potential outcomes θ_Y. A common assumption that is not always innocuous is that the parameters entering the marginal distribution of the covariates are distinct from those entering the conditional distribution of the potential outcomes given the covariates, i.e., our joint distribution can be written as $f(Y(0), Y(1), X|\theta_{Y|X}, \theta_X)$.