# Subclassification on the Propensity Score

Lorenzo Lagos
March 7, 2017

This chapter goes through the details of different ways of applying the propensity score. (Chapter 18 does something similar for matching). As we know, the propensity score is simply the conditional probability of receiving treatment given the observed covariates. The effectiveness of removing biases through this approach relies on unconfoundedness (which is still quite a strong assumption in any observational study). Imbens and Rubin don't seem to discuss whether the model that we use to determine the balancing score matters much for our results. Using a linear probability model or a logit may be important and it is unclear what sort of scientific knowledge the researcher can use to make such a decision.

What the authors refer to subclassification is simply splitting the sample into groups (or blocks) based on their propensity score, calculating the average treatment effect in each block, and then estimating the overall treatment effect by weighing the block-specific treatment effects. The choice of blocks and their boundaries seems rather ad-hoc, but Imbens and Rubin have interesting algorithms to provide these. They also show that the bias from such subclassification is attributed to the weighted average of the within-block biases. This can arise from failure of unconfoundedness (assumed away) or because propensity scores are only approximately constant within each block. The latter can be adjusted with regression adjustment or model-based imputation (which the authors explained in detail). However, my main comment here (and with everything we have done so far in class) is that we rely on unconfoundedness—a thing that can neither be proved nor disproved. Wouldn't it be much better to rely on quasi-experimental variation rather than modeling assumptions for identifying causal effects? That is, shouldn't we hunt for events that introduce reasonable exogenous variation in our observational setting and exploit?

In any case, Imbens and Rubin end their chapter with a discussion about inverse propensity score weighted regressions. The authors prefer subclassification because weighing by the inverse of the propensity score can generate considerable bias due to the fact that the estimated score is a noisy estimator. It still is unclear to me why the noisiness of the propensity score is not as much of a concern for subclassification. But in any case, I think any causal inference using these methods in observational studies has low credibility due to the fact that it is model based.