Matching and Observational Studies

Alan Liang

May 2018

1 Observational Studies

1.1 Motivation

In the last few lectures, we focused on conducting randomized experiments and dealing with some of its nuances such as noncompliance, attrition, or sequential monitoring. However, in the real world it may become very costly or impossible to conduct randomized experiments. In this case, we resort to observational studies. Similar to randomized experiments, observational studies may also record a treatment variable and the outcome, but generally the study is conducted without randomization: sometimes randomization is practically infeasible, especially for large scale-studies
As data scientists, we would still like to conclude causational relationship, but how can we approximate the observational study as a randomized experiment?

1.2 Some Definitions

In all observational studies we will continue to see some form of treatment variable T_i being applied to a subset of the sample. With the treatment being applied, any study will also measure some outcome variable Y_i. However, as observational studies are done in the real world, it cannot escape possible confounding variables.
A covariate, denoted X_i, is a potential confounding variable that may also affect the outcome variable and/or treatment. In the real world, there could be countless covariates out there which we may not even be aware of. These next few notes will go over ways we can account for some of them.
Consider an investigation on the effects of Yelp display ratings on restaurant popularity. Although a display rating may be correlated with being popular, it does not set up causation. In fact, many confounding covariates contribute to this: perhaps a restaurant may have a good reputation, which causes both its popularity and high display rating on Yelp.

2 Matching

Matching is the most basic and probably most intuitive way to deal with possible covariates. At its core, matching simply compares a treated unit to similar non-treated unit(s) in order to **impute** the effect of the treatment. The large assumption matching makes is that if 2 units are similar in the feature(s) we observe and choose to match on, then they will be similar in all covariates and aspects, even the unobserved potential covariates. This is called the **selection on observations assumption**, which states that after we account for the covariates, the treatment group is as good as randomly selected and assigned, which would in turn mean that the treatment is independent of the potential outcome. In probability terms, we are assuming that a unit who receives the treatment will have the same potential outcomes as a similar unit on X_i who does not receive the treatment:

$$E[Y_{i0}|D_i = 1, X_i] = E[Y_{i0}|D_i = 0, X_i]$$

The conclusion is that we have essentially controlled for all possible covariates. It's a naive assumption, but it's the best we got so far.

2.1 Ways to do Matching

We can match units based on one or more different variables. However, it is important to note that as more features and covariates are being considered to match for similarity, the "curse of dimensionality" makes an exact matching over many variables more and more infeasible.
Say you are doing an experiment on the effect of training on wages, but suspect that age may be a confounding variable: perhaps older people are also more likely to receive higher wages. To do matching, for every unit that received the training, you would match it with a similar unit in terms of age who did not receive the training and then compare the differences.
We can do matching in a few ways:

- Out of all equally similar units, we can pick a random unit and match its outcome variable to the treatment unit's outcome variable
- Out of all equally similar units, we can take the units' average outcome variable and match that to the treatment's outcome variable.
- If there are no exactly similar units, a solution could be to match to the nearest neighbor(s). Here, distance and hence similarity is up to you to define: perhaps you can assume an n-dimensional space and use Euclidean distance (think back to $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 \cdots}$), or perhaps you may weigh certain covariates more heavily and define your own metric. The bottom line is that variables must be similar, especially with respect to the potential covariates.